Utrecht University

FAME

Feedback & Assessment in Mathematics Education

FAME 1 Proceedings

Feedback & Assessment in Mathematics Education

5-7 June, 2024 - Utrecht, The Netherlands

Edited by:

Paola Iannone, Filip Moons, Christina Drüke-Noe, Eirini Geraniou, Francesca Morselli, Katrin Klingbeil, Michiel Veldhuis and Shai Olsher

erme

european society for research in mathematics education

14th ERME Topic Conference (ETC14)

The International Programme Committee (IPC)

Chair of the IPC:

Paola Iannone (United Kingdom)

Members:

Christina Drüke-Noe (Germany)
Eirini Geraniou (United Kingdom)
Filip Moons (Belgium)
Francesca Morselli (Italy)
Katrin Klingbeil (Germany)
Michiel Veldhuis (The Nettherlands)
Shai Olsher (Israel)

The Local Organising Committee (LOC)

Filip Moons (Belgium)
Michiel Veldhuis (The Netherlands)

Conference webpage:

https://www.uu.nl/fame

Editors: Paola Iannone, Filip Moons, Christina Drüke-Noe, Eirini Geraniou, Francesca Morselli, Katrin Klingbeil, Michiel Veldhuis and Shai Olsher
Publisher: Utrecht University
Place: Utrecht, The Netherlands
Year: 2024
ISBN: 978-90-393-7788-8
DOI: 10.5281/zenodo.14231455

Recommended citation for the proceedings

Iaonnone, P., Moons, F., Drüke-Noe, C., Geraniou, E., Morselli, F., Klingbeil, K., Veldhuis, M., Olsher, S. (Eds.). (2024). *Proceedings of FAME 1 – Feedback & Assessment in Mathematics Education (ETC 14), 5-7 June 2024, Utrecht (The Netherlands)*. Utrecht University and ERME. https://doi.org/10.5281/zenodo.14231455

Recommended citation for single entries in the proceedings

Authors (2024). Title of paper/poster. In P. Iaonnone, F. Moons, et al. (Eds.). *Proceedings of FAME 1 – Feedback & Assessment in Mathematics Education (ETC 14), 5-7 June 2024, Utrecht (The Netherlands)* (pp. xx-yy). Utrecht University and ERME. https://doi.org/10.5281/zenodo.14231455

Introduction

FAME1 – A first ERME topic conference on Feedback and Assessment in Mathematics Education

Assessment is pervasive in the teaching and learning of mathematics at any educational level though often its impact on the students' and teachers' experiences is under-estimated. During the work of the TWG21 - Assessment in Mathematics Education which will be led by Francesca Morselli at CERME14 in February 2025 – we realised that research on the assessment of mathematics is present in many TWGs at CERME. Therefore, we decided to organise the ERME topic conference FAME - Feedback and Assessment in Mathematics Education - to be a forum where all those interested in research related to the assessment of mathematics could meet. The conference took place in Utrecht between the 5th and 7th of June 2024. The call for papers highlighted three themes:

Theme 1: Formative feedback on mathematics tasks and its impact on students' experiences of learning mathematics: students' and teachers' perspectives. Within this theme we received papers related to the effect and impact of formative feedback on students' mathematics learning as well as on teachers' perspectives, at any educational level.

Theme 2: Resources for the (formative or summative) assessment of mathematics and design of tasks for assessing specific mathematics topics (as for example in the work of TWG22, TWGs 1 to 6 for specific mathematics topics). The issue of task-design for the assessment of mathematics cannot be overlooked and this theme comprised contributions that focused on task design for the assessment of specific topics (e.g., geometry or proof).

Theme 3: Teachers' and students' experiences with technology in/for the (summative and/or formative) assessment of mathematics (related to the work of TWG15/TWG16). Technology has become present in the teaching of mathematics but specifically in assessment and feedback. This theme collected contributions which addressed the use of technology for both summative and formative assessment.

At the conference 43 papers and 7 posters were presented. Topics were spread along the three themes, but we also had papers that spanned more than one theme. The papers and posters presented are included in these proceedings.

We had the pleasure to have two invited plenary talks, the first by Michal Ayalon (University of Haifa, IL) who spoke about the benefits of encouraging feedback literacy with trainee teachers and the second by Juuso Nieminen (Hong Kong University, HK) who spoke about student agency in assessment practices. We thank the plenary speakers for their interventions and their papers are also included in these proceedings.

The plenary panel was coordinated by Fracesca Morselli (University of Genova, Italy) and included presentations on the theme of the impact of summative assessment at the secondary level in mathematics on students' perceptions and attitudes. We thank out panel speakers, Paul Drijvers (Utrecht University, NL), Alice Lemmo (Universita' dell'Aquila, IT) and Hans-Georg Weigand (University of Würzburg, DE) for their contributions and for the very stimulating debate that followed.

We were of course delighted to see many young researchers at the conference, and for them we organised an event on the morning of the 4th June, before the official start of the conference, for networking and for discussing the process of becoming a researches in mathematics education. Thank you to Katrin Klingbeil, Filip Moons, Eirini Geraniou and Christina Drüke-Noe for having organised the morning activities.

I want to close this brief introduction of the Proceedings of FAME by thanking the IPC and the local organizing committee – without them nothing would have been possible! They are

International Programme Committee

Christina Drüke-Noe	Pädagogische Hochschule Weingarten (DE)
Eirini Geraniou	University College London (UK)
Filip Moons	Utrecht University (NL)
Francesca Morselli	Universita' di Genova (IT)
Katrin Klingbeil	University of Duisburg-Essen (DE)
Michiel Veldhuis	Utrecht University (NL)
Paola Iannone	University of Edinburgh (UK)
Shai Olsher	University of Haifa (IL)

Local Organizing Committee

Filip Moons	Utrecht University (NL)
Michiel Veldhuis	Utrecht University (NL)

Finally – we are already planning for FAME2 which will be held in Budapest in 2026 – I hope to see you all there!

Paola Iannone
President of the IPC of FAME1

The University of Edinburgh
United Kingdom
paola.iannone@ed.ac.uk

Table of Contents

Plenary talks & Panel discussion

Papers

Posters

Plenary talks & Panel discussion

Giving and receiving feedback on mathematics tasks: Insights gleaned from exploring teachers' and students' perspectives

Michal Ayalon

University of Haifa, Israel; mayalon@edu.haifa.ac.il

This paper explores the opportunities and challenges in cultivating feedback literacy among mathematics teachers and students through formative assessment practices. Drawing on findings from three distinct studies—Study A, which investigates pre-service teachers' engagement with feedback; Study B, which delves into students' engagement with feedback; and Study C, which examines the engagement of both teachers and students with feedback—this research provides a comprehensive analysis across varied educational contexts. The results demonstrate a significant enhancement in feedback literacy for both teachers and students, marked by a deeper understanding and more effective utilization of feedback. However, the findings also highlight the complexities and obstacles in implementing feedback practices, particularly within the context of mathematics education. The paper concludes by identifying key areas for further research, aiming to advance the effective integration of feedback literacy in educational practice.

1. Introduction

Feedback plays a crucial role in the learning process (Hattie & Timperley, 2007). It not only enhances understanding but also encourages metacognitive skills such as self-regulation and reflection (Shute, 2008). Despite its importance, there is significant variability in how feedback is provided, received, and utilized in the classroom. This paper aims to share findings from research focused on feedback literacy among mathematics teachers and students. It also discusses the feasibility of implementing effective feedback processes in the mathematics classroom.

2. Literature Review

2.1 Feedback

According to the Oxford Dictionary, feedback is defined as "advice, criticism, or information about how good or useful something or somebody's work is." When examining educational research on feedback, opinions vary on several key aspects, such as who should provide feedback (teachers or peers), how and when feedback should be delivered (e.g., written, spoken, planned, unplanned), and what the focus of feedback should be (e.g., whether it should address the correctness of work or focus on the processes and strategies underlying the task) (e.g., Carless & Winstone, 2020; Hattie & Timperley, 2007; Rakoczy et al., 2019; Sadler, 1989). Despite these differing viewpoints, there is some consensus on certain aspects of feedback that are effective for student learning, such as focusing on thinking processes and self-regulation, providing clear and specific feedback, and teaching students about assessment criteria (Fujita et al., 2018; Smit et al., 2023; van der Kleij, 2019; Wiliam & Thompson, 2008).

Traditionally, feedback has been viewed as a one-way transmission of information from teacher to student, aimed at correcting errors and guiding future performance (Hattie & Timperley, 2007). However, more recent perspectives emphasize the role of feedback as a dialogic process that involves both the giver and the receiver. This shift towards a more interactive and student-centered approach to feedback recognizes that learners are active participants in interpreting and using feedback to enhance their learning (Boud & Molloy, 2013). According to Carless and Boud (2018), feedback should be understood as "a process through which learners make sense of information from various sources and use it to enhance their work or learning strategies." (Carless & Boud, 2018, p, 1315). This definition goes beyond notions that feedback is principally about teachers informing students about strengths, weaknesses and how to improve. It highlights the centrality of the student role in sense-making and using comments to improve subsequent work.

Research indicates that student engagement in feedback enhances understanding, encourages the development of metacognitive skills such as self-regulation and reflection, boosts students' confidence, reduces math anxiety, contributes to a positive and supportive classroom culture, and ultimately improves learning outcomes. (Adarkwah, 2021; Carless & Winstone, 2023; Hattie & Timperley, 2007; Mahfoodh, 2017; Shute, 2008; van der Kleij et al., 2019). These studies highlight the importance of implementing and researching feedback processes in mathematics classrooms.

2.2 Student and teacher feedback literacy

A central concept in the work presented here is feedback literacy. In terms of students' feedback literacy, Carless and Boud (2018) define it as "The understandings, capacities and dispositions needed to make sense of information and use it to enhance work or learning strategies." (Carless & Boud (2018, p, 1316). They identify four key features of feedback literacy: first, students must appreciate their active role in the feedback process, recognizing that they are not merely passive recipients but active participants who need to engage with and act upon the feedback they receive, viewing it as a learning tool rather than just a judgment of performance. Second, feedback literacy involves the ability to make sound judgments about the quality of work, enabling students to assess both their own and others' work against established criteria, which supports self-regulation and the ability to independently enhance their performance. Third, it encompasses the ability to manage affect in constructive ways, as feedback can sometimes evoke strong emotional reactions, particularly when perceived as negative or critical; thus, feedback literacy helps students manage these emotions positively and see feedback as an opportunity for growth rather than a source of discouragement. Finally, feedback literacy is about taking action, where students must be able to convert feedback into actionable steps for improving their learning, which requires a clear understanding of the feedback and how to apply it in future tasks (Carless & Boud, 2018).

Teachers play a crucial role in designing and delivering feedback processes that support student learning in general, and students feedback literacy in particular (Carless & Winstone, 2023). Carless and Winstone (2023) define teachers' feedback literacy as "The knowledge, expertise and dispositions to design feedback processes in ways which enable student uptake of feedback and seed the development of student feedback literacy." They identify three dimensions of teacher feedback literacy: the design dimension, which focuses on the teacher's ability to create feedback processes that encourage student engagement and foster the development of student feedback literacy by designing tasks that generate meaningful feedback and rubrics that clearly define success criteria (Brookhart, 2013); the relational dimension, which emphasizes the significance of the interpersonal relationship between teacher and student, requiring teachers to be supportive and sensitive in their delivery of feedback, thereby influencing students' motivation and self-esteem (Wiliam, 2011) and ensuring that students feel comfortable and confident in receiving and applying feedback; and the pragmatic dimension, which addresses the practical challenges of delivering feedback within the constraints of the classroom and broader educational context, where teachers must balance the need for individualized feedback with the limitations posed by time constraints and large class sizes (Looney et al., 2018). Carless and Winstone (2023) also note that feedback literacy is closely tied to the particular discipline in which one works.

2.3 Feedback in practice

Research on feedback has shown that many students do not act on the feedback they receive, often lack effective strategies to utilize it, and struggle to recognize what constitutes a high-quality solution (Black and Wiliam 2009; Jönsson, 2013; van Gennip et al., 2010). Both students and teachers frequently view feedback primarily as information on strengths, weaknesses, and ways to improve student work (Little et al., 2024). This narrow perspective can limit the effectiveness of feedback as a tool for deeper learning and development. However, there is evidence that targeted interventions can successfully enhance features of feedback literacy (Ketonen et al., 2020; Little et al., 2024). Despite these promising findings, most studies in this area have been conducted at the university level (Little et al., 2024), with limited attention given to feedback practices in school mathematics. This gap is particularly concerning because mathematics often involves complex problem-solving and abstract reasoning, making effective feedback both more challenging to provide and more critical for student success (Schoenfeld, 2007; Teledahl, 2017). The research presented here addresses these gaps by focusing on the opportunities and challenges in developing feedback literacy among mathematics teachers and students, with a particular emphasis on the school mathematics context.

3. Three studies

This paper draws on three studies focusing on feedback literacy in different educational contexts: Study A, which examines pre-service teachers' (PSTs) engagement with feedback; Study B, which explores students' engagement with feedback; and Study C, which investigates teachers' and students' engagement with feedback. Findings from Study A have been published in Ayalon and Wilkie (2020). Studies B and C were conducted with a group of MA students and are currently in the process of being written. The common goal across these studies is to investigate whether and how experiencing formative assessment processes can support the development of feedback literacy among teachers and students. The research design for each study involved cycles of working on math tasks, generating and assessing feedback using rubrics, refining solutions and rubrics, and reflecting on the process.

The studies employ Looney et al.'s (2018) framework for Teacher Assessment Identity to examine various aspects of feedback literacy in both teachers' and students' reflections on their experiences with feedback. Looney et al. (2018) conceptualized teachers' assessment work as being deeply intertwined with their professional identity, beliefs about assessment, disposition towards enacting assessment, and their perceptions of their role as assessors. The framework they proposed encompasses five interconnected aspects—articulated in the first person to emphasize their self-reflective nature: *I know*, *My role*, *I believe*, *I am confident* (self-efficacy), and *I feel*. In our analysis of teachers' and students' reflections across three studies, we applied these categories through iterative processes of data sorting, continual comparison between the data and the evolving categories, and cross-category analysis. The findings are presented separately for each study below.

3.1 Study A. Pre-service teachers' engagement with feedback

Goal. Study A focused on investigating whether and how the experience of approximations of formative assessment practices could support the development of feedback literacy among pre-service teachers (PSTs).

Population. The study involved 97 pre-service teachers across two cohorts.

Research design. The research design incorporated a cycle that included small group work, the creation of rubrics, the assessment of student solutions, the refinement of these rubrics, and individual reflection (Figure 1).

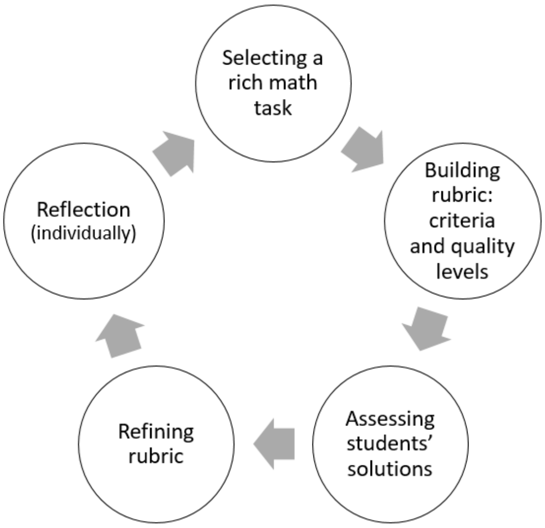

Selecting a rich math task
Building rubric: criteria and quality levels
Assessing students' solutions
Refining rubric
Reflection (individually)

Figure 1. Study A cycle

Data collection included (1) PSTs' written responses, including (i) initial and revised assessment criteria and accompanying explanations, and (ii) reflections on their noticed differences between their initial and final assessment criteria and their learning about assessment practices, and (2) group interview that focused on the PSTs' experiences, their perceived strengths and difficulties, and their learning gains.

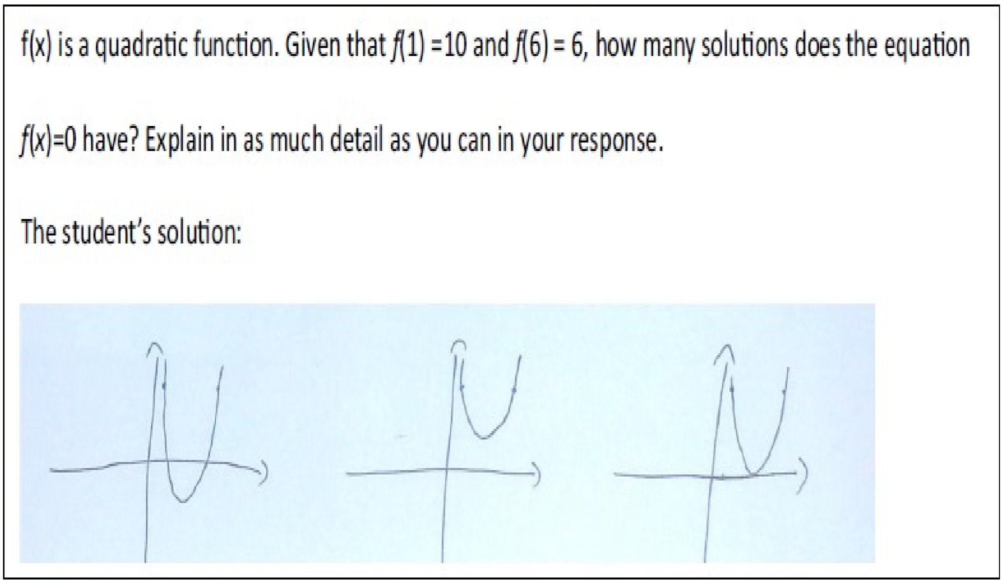
Data analysis included (1) iterative and comparative process for generating categories for the assessment criteria created by the PSTs, and (2) Drawing on Looney et al.'s (2018) categories for analyzing the PSTs' reflections: feelings, beliefs, self-efficacy, role enactment, and knowledge.

Findings. Overall, the PSTs reflected that the formative assessment (FA) processes significantly enhanced their understanding of feedback. They reported gaining valuable skills in building rubrics, identifying strengths and weaknesses, and proposing improvements. Additionally, there was evidence of a shift in their perception of their role in the feedback process, moving from a narrow view to recognizing the student as an integral part of the feedback cycle. Below are the main emerging themes, some of which are accompanied with examples.

Theme 1. *Awareness of the need for clear feedback and for students to understand the criteria.*

Theme 2. *Uncertainty in applying some assessment criteria, considering the student's perspective.* For example, one of the PSTs wrote:

> When I tried to assess the students' responses for their *quality of communication*, I felt confused [I feel]… there were some cases in which the solution 'spoke for itself' and I did not need further explanation to understand its underlying thinking. So, should I give a lower assessment for not including an explanation? [I am not confident]. For example, a student drew three parabolas to answer the functions task [Figure 2]. I like the solution; it reflects understanding, but I still wonder, should I also require a descriptive explanation from him? Is the solution perfect regarding communication [I am not confident]? I am thinking of the student, I prefer not to burden him with unnecessary work and then step back from the feedback process [I believe]." (PST3)



$f(x)$ is a quadratic function. Given that $f(1) = 10$ and $f(6) = 6$, how many solutions does the equation $f(x) = 0$ have? Explain in as much detail as you can in your response.

The student's solution:

Figure 2. The functions task

In this example, the PST considered the student's perspective, aiming to avoid unnecessary frustration.

Theme 3. *Awareness of the limitations of their own interpretations of students' solutions.* For example, one of the PSTs wrote:

> What was difficult for me sometimes was being sure about my understanding of a student's way of thinking based on reading his answer [I am not confident]. The task is open-ended, and the solution is not obvious in its presentation or reading, which makes it hard to assess. Indeed, there was a case where my peer and I interpreted a student's solution differently. The student presented her solution but obtained the wrong answer. The explanation was very vague. I thought she had made a calculation error, but my colleague thought the student had misunderstood the task questions. So, to ensure the student receives a fair score and constructive comments, it's beneficial not to depend solely on my perspective [I believe]. Engaging the student for additional

> explanations about their answer can enhance my understanding of their knowledge, moving closer to an accurate assessment. [My role]" (PST26)

In this example, the PST considered the student's perspective by prioritizing fairness, providing productive feedback, and involving the student in the process.

Theme 4. *Understanding the importance of providing supportive and motivational feedback.*

Theme 5. *Recognition of the importance of engaging students in exploring their solutions through feedback.* For example, one of the PSTs wrote:

> My colleague and I used the Truck task [finding the maximum number of pallets that a truck can carry under a bridge] with 7th-grade students and chose these criteria to assess their solutions: the model used, manipulations, inferring, communication, and creativity. When I first evaluated their work, I assessed each student's quality level for each criterion and provided accompanying explanations. However, in discussions with my colleague, we reflected on how engaged the students were with the task and that we must keep them engaged with the feedback so they don't just glance at their assessments and dismiss them [My role]. So, instead of directly telling students what was good or bad, we activated them in the feedback process. For instance, some students provided valuable solutions but overlooked some of the constraints given in the situation. Here, my feedback was related to the "Inferring" criterion. I wrote: "Dan, the man in the task, is very worried. Try to convince him that your solution perfectly meets his requirements." This exemplifies how we engaged students in the feedback to motivate them to improve [I believe]." (PST 45)

In this example, the PST considered the student's perspective, aiming to engage them in exploring the solution with the feedback.

Summary. The summary of findings indicates a notable shift among PSTs toward recognizing the student as an integral part of the feedback process. This shift aligns with the dimensions of feedback literacy as outlined by Carless and Winstone (2023), particularly in the areas of design for uptake and the relational dimension. The change is evident across various aspects, including knowledge, feelings, beliefs, roles, and confidence, as described by Looney et al. (2018). However, some potential clashes emerged, such as uncertainty in applying criteria ("I know the criteria, but am unsure if I need to use them"), difficulty in understanding their role ("I see it as my role, but I don't know how to apply it"), and a lack of confidence despite their beliefs ("I believe, but I am not confident"). The main explanation for this shift, from the PSTs' perspective, was the opportunity to encounter and discuss different perspectives on student responses, which emphasized the need for a more nuanced, open, and responsive approach to providing feedback.

3.2 Study B. Students' engagement with feedback

Goal. Study B focused on investigating if and how a process of experiencing formative assessment practice might support students' development of feedback literacy.

Population. The study involved 147 10th-grade students.

Research design. The research design incorporated a cycle that involved small groups solving rich mathematical tasks, building rubrics, assessing solutions, refining the rubrics, and individually reflecting on the process (Figure 3).

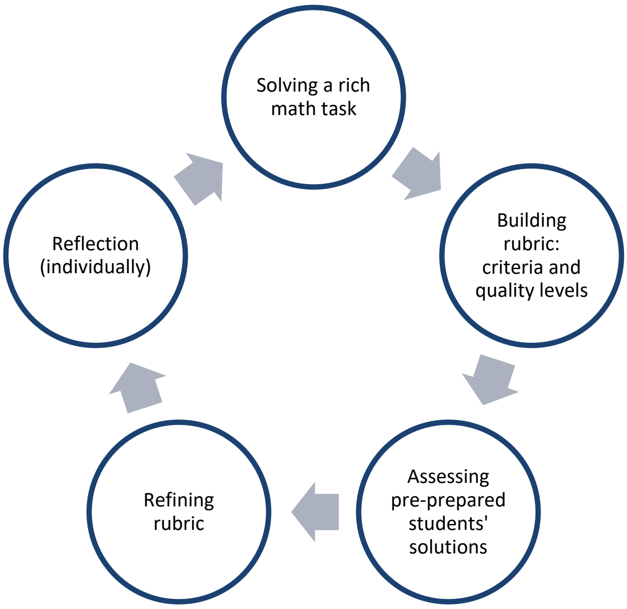

Solving a rich math task
Building rubric: criteria and quality levels
Assessing pre-prepared students' solutions
Refining rubric
Reflection (individually)

Figure 3. Study B cycle

Data collection included (1) students' written responses, which comprised (i) task attempts, (ii) initial and revised assessment criteria with accompanying explanations, and (iii) final reflections on the differences they observed between their initial and final assessment criteria and their learning about assessment practices, and (2) Individual interviews that focused on the students' experiences, perceived strengths and difficulties, and learning gains.

Data analysis included (1) an iterative and comparative process for generating categories for the assessment criteria created by the PSTs and examining the distribution of these criteria, and (2) utilizing Looney et al.'s (2018) categories to analyze the students' reflections, focusing on feelings, beliefs, self-efficacy, role enactment, and knowledge.

Findings. Data analysis revealed four main emerging themes, which are presented below, with some accompanied by examples.

Theme 1: *Increased, but still restricted, knowledge about using criteria for assessing solutions*. For example, one of the PSTs wrote:

> At the beginning I did not fully understand the assessment criteria. But while using them to assess one more solution and another different solution etc., I better understood what they mean. Some criteria were more difficult for me to use. For example, I wasn't sure whether to give preference to a certain solution strategy over another, such as algebraic, visual, verbal. I also wasn't sure how to determine whether a solution is original or not – if I did not think about it? [I know]" (S18)

Theme 2: *A shift in perspectives related to goals and the importance of feedback*. Some PSTs emphasized that feedback should include the reasoning behind it to help students understand why and how to proceed. For example,

> I tried to get into the heads of the students for whom I wrote the feedback and think about how best I could support them [My role]. I wanted the feedback to, like, connect with the student. I understand now that feedback shouldn't simply highlight errors and their corrections but also provide reasoning behind the mistakes and guide the student towards another attempt [I know]." (S18)

Additionally, some PSTs highlighted that feedback criteria can teach us about the nature of doing mathematics. For example,

> Assessing the students' solutions through the rubric criteria made me realize that the feedback highlights various mathematical aspects, teaching us that there's much more to mathematics than just the final answer. For example, consider whether our mathematical strategy is good or whether it is worth trying an alternative solution, perhaps using a graph or algebra. Or if we used the data appropriately… I understand now that the feedback provides us with cues to consider in advance when approaching a task [I know]." (S43)

Theme 3: Experience anxiety initially but ultimately felt satisfaction from generating feedback. For example,

> Writing the feedback is challenging. At first, I felt nervous [I feel] because I thought I couldn't do it [My confidence]. However, through various experiences and discussions with friends, I felt more confident [My confidence]… I still feel a bit anxious, especially because sometimes I am not sure that I understood the other student's thinking (I know). But it also feels worthwhile. It is like "getting out" of myself and looking at others' work, and then, when I "return" to look at my work, I can better assess it [I know]. It feels good [I feel]." (S52)

Theme 4: Awareness of their role as students in acting on feedback. For example,

> As I wrote feedback for the students, I realized that simply reading the feedback isn't sufficient for the student to learn and improve. He should go over his solution and improve it with my feedback. I realized that when my teacher gives me feedback, it's the same. I really need to rethink my solution with the feedback. It's up to me if I want to learn from it [My role]". (S3)

Summary. The summary indicates a shift among the participants toward recognizing themselves as integral parts of the feedback process. This shift aligns with key aspects of feedback literacy, as outlined by Carless and Boud (2018), including appreciating feedback, making judgments, and managing affect. These changes were evident across various dimensions, such as knowledge, feelings, beliefs, roles, and confidence, as described by Looney et al. (2018), though potential clashes emerged. For instance, participants expressed sentiments like, "I see it as my role, but I feel anxious about it," and "I know the criteria, but I am unsure how to apply them." The main explanations for this shift include the experience of assessing imaginary students' solutions, which varied across different criteria and helped ease pressure, as well as the group's collective discussion and negotiation on the feedback.

3.3 Study C. Teachers' and students' engagement with feedback

Goal. Study B focused on investigating if and how a process of experiencing formative assessment practice might support teachers and students' development of feedback literacy.

Population. The study involved 20 secondary teachers and 200 students.

Research design. The teachers participated in a university course focusing on formative assessment, consisting 12 sessions. In the first session, the teachers discussed how feedback is used in their classrooms. Following this session, they administrated a feedback questionnaire to their students, focusing on Looney et al.'s (2018) categories to of feelings, beliefs, self-efficacy, role enactment, and knowledge. In the next session, the teachers shared their students' responses and concluded that, generally, the students did not appreciate feedback. As a result, they decided to plan an intervention aimed at engaging students more actively in the feedback process. The following two sessions were dedicated to learning about formative assessment and feedback, as well as planning the intervention. The final eight sessions focused on three cycles of experiencing feedback. Each cycle involved designing tasks and rubrics, implementing them in the classrooms, and sharing experiences to inform further decisions (Figure 4).

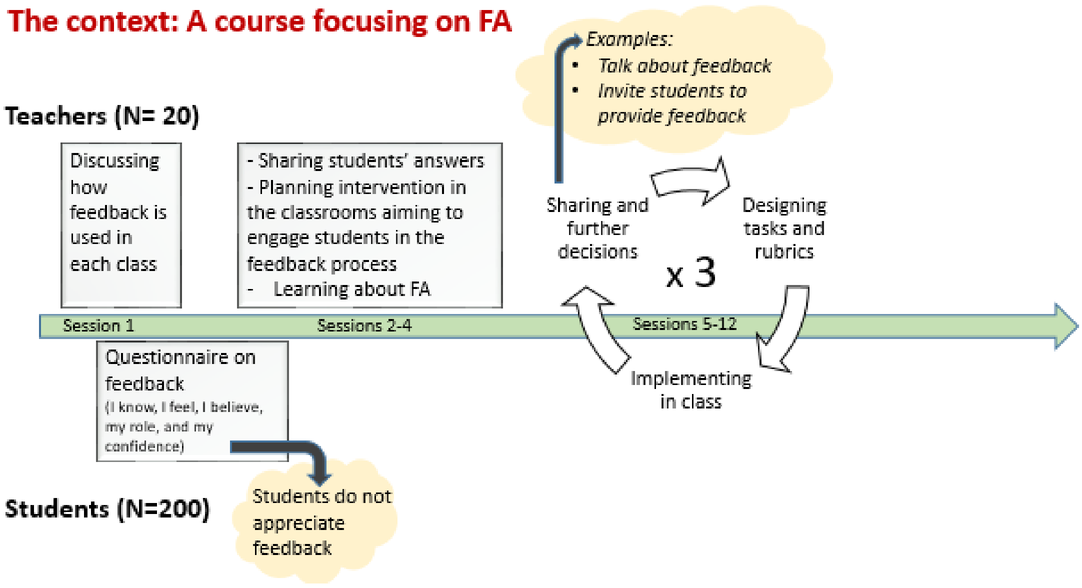

The context: A course focusing on FA
Teachers (N= 20)
Discussing how feedback is used in each class
- Sharing students' answers
- Planning intervention in the classrooms aiming to engage students in the feedback process
- Learning about FA
Session 1
Sessions 2-4
Sessions 5-12
Questionnaire on feedback
(I know, I feel, I believe, my role, and my confidence)
Students (N=200)
Students do not appreciate feedback
Examples:
• Talk about feedback
• Invite students to provide feedback
Sharing and further decisions
Designing tasks and rubrics
x 3
Implementing in class

Figure 4. Study C design

In this paper, I focus on the students' perspective. The students participated in three cycles of experiencing feedback, which included solving a rich mathematics task, receiving feedback from their teacher, refining their solutions, providing "feedback on the teacher's feedback, and class discussion (Figure 5).

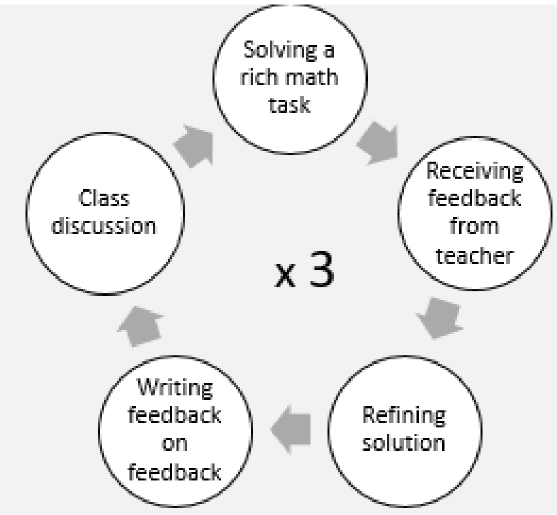


Figure 5. Study C cycle (students)

Data collection included pre-questionnaire and post-questionnaire (administrated after the three cycles) on feedback, focusing on Looney et al.'s (2018) categories to of feelings, beliefs, self-efficacy, role enactment, and knowledge.

Data analysis utilized Looney et al.'s (2018) categories to analyze the students' responses to the pre- and post- questionnaires on feedback.

Findings. Data analysis revealed four main emerging themes that reflect shifts in students' perspectives to feedback from the pre-questionnaire to the post-questionnaire. These themes are presented below, accompanied by examples from the responses of the same student, Lia, to illustrate a main profile identified among the students.

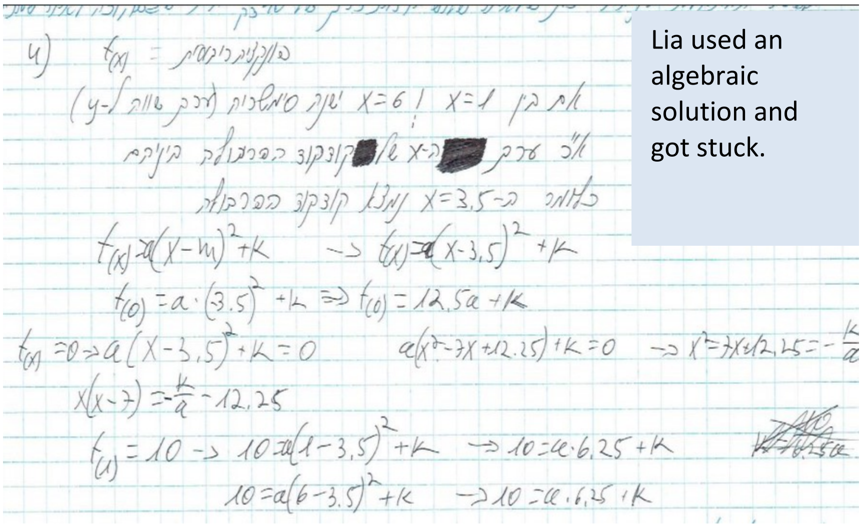
Theme 1: *A shift in perspectives related to goals and the importance of feedback.* For example, in the pre-questionnaire, Lia wrote:

> Feedback clarifies why points were deducted [I know]. The teacher provides feedback so we can correct our mistakes. [My role].

In the post-questionnaire, Lia wrote:

> The assessment criteria in the feedback are crucial for highlighting what's important in the task, and what can aid in finding a successful solution [I know]. I believe that the strategies I learn from the feedback will help me in solving other problems in the future [I believe] … For example, the feedback about the strategy I used in my solution to the functions task taught me not to rush into the first strategy that comes to mind, but to pause and consider whether the strategy is truly appropriate for the task [I know].

In her response, Lia refers to the feedback she received from her teacher to her solution to the functions task [Figure 2]. Figure 6 presents Lia's solution in Hebrew. Lia used an algebraic approach to solve the solution and got stuck.



4) $f_{(x)}$ = פונקציה ריבועית

אם בין $X=1$ ו $X=6$ ישנה סימטריות (ערך שווה ל-y)

א״כ ערך [redacted] ה-x של [redacted] קודקוד הפרבולה ביניהם

כלומר ב-$X=3.5$ נמצא קודקוד הפרבולה

$f_{(x)}=a(x-h)^2+k \rightarrow f_{(x)}=a(x-3.5)^2+k$

$f_{(0)}=a\cdot(3.5)^2+k \Rightarrow f_{(0)}=12.5a+k$

$f_{(x)}=0 \rightarrow a(x-3.5)^2+k=0 \quad a(x^2-7x+12.25)+k=0 \rightarrow x^2-7x+12.25=-\frac{k}{a}$

$x(x-7)=-\frac{k}{a}-12.25$

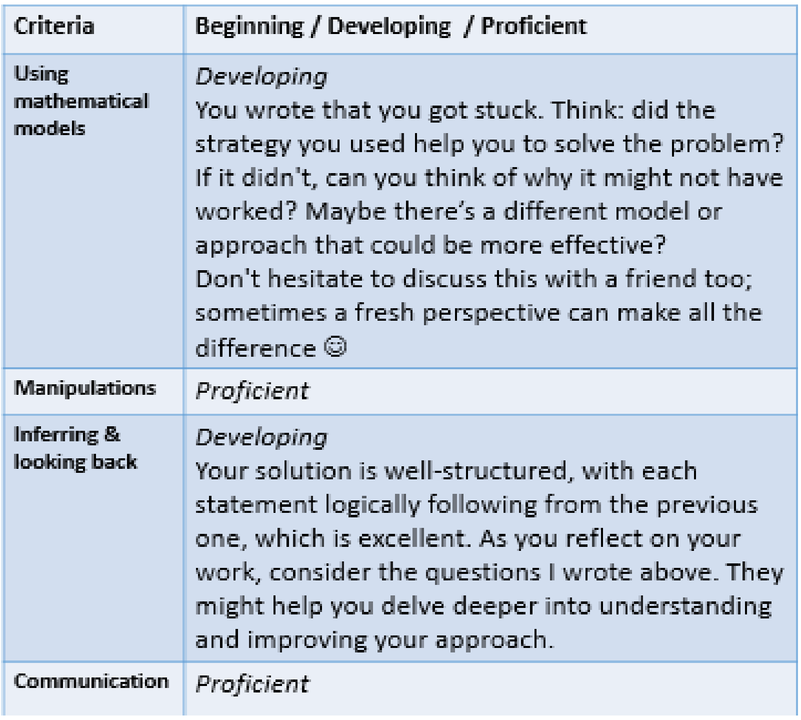
$f_{(1)}=10 \rightarrow 10=a(1-3.5)^2+k \rightarrow 10=a\cdot6.25+k$

$10=a(6-3.5)^2+k \rightarrow 10=a\cdot6.25+k$

Lia used an algebraic solution and got stuck.

Figure 6. Lia’s solution to the functions task

Figure 7 presents the teacher’s feedback, which was structured using a rubric with four criteria: using mathematical models, manipulations, inferring and looking back, and communication. For each criterion, the teacher provided an assessment of Beginning, Developing, or Proficient, accompanied by an explanation.



Criteria	**Beginning / Developing / Proficient**
Using mathematical models	*Developing* You wrote that you got stuck. Think: did the strategy you used help you to solve the problem? If it didn't, can you think of why it might not have worked? Maybe there’s a different model or approach that could be more effective? Don't hesitate to discuss this with a friend too; sometimes a fresh perspective can make all the difference ☺
Manipulations	*Proficient*
Inferring & looking back	*Developing* Your solution is well-structured, with each statement logically following from the previous one, which is excellent. As you reflect on your work, consider the questions I wrote above. They might help you delve deeper into understanding and improving your approach.
Communication	*Proficient*

Figure 7. The teacher’s feedback to Lia’s solution

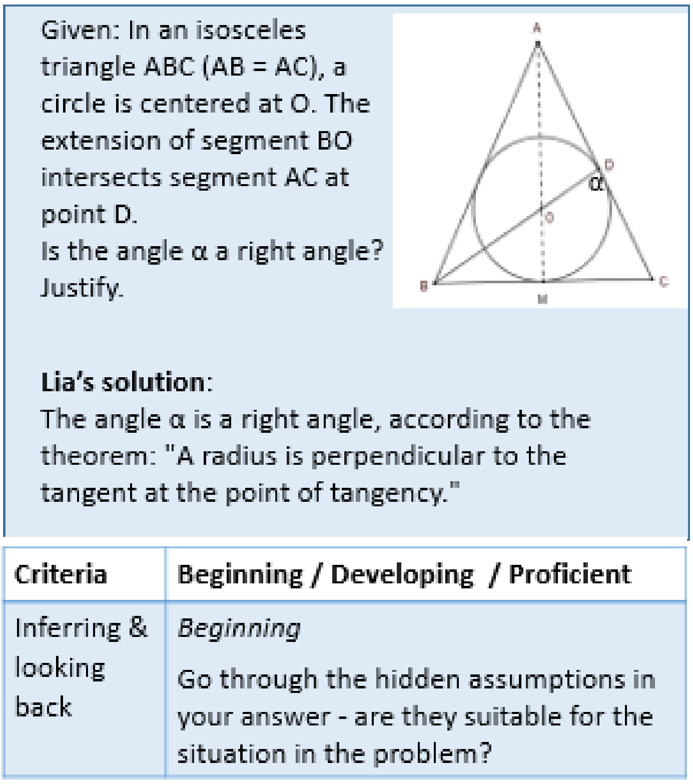
As evidenced by Lia’s responses to the pre- and post-questionnaires presented above, there is a shift in her perception of feedback from merely correcting mistakes to recognizing it as a valuable source of learning.

Theme 2: *Increased understanding of feedback.* In the post-questionnaire, Lia wrote:

> Over time, I've realized that my teacher gives me feedback by asking questions and making comments to help me figure out my solutions and see what I did right or wrong. But sometimes, it

> is tough for me to grasp what the teacher is thinking. Like, for instance, when I got feedback on a geometry problem about making inferences, the teacher asked me a question to think about, but I just couldn't understand what she was getting at. It's like she had a clear idea, but I couldn't quite see things from her perspective. [I know]

In her response, Lia refers to the feedback she received from her teacher to her solution to a geometry task [Figure 8].



Given: In an isosceles triangle ABC (AB = AC), a circle is centered at O. The extension of segment BO intersects segment AC at point D.
Is the angle α a right angle? Justify.

Lia's solution:
The angle α is a right angle, according to the theorem: "A radius is perpendicular to the tangent at the point of tangency."

Criteria	**Beginning / Developing / Proficient**
Inferring & looking back	*Beginning* Go through the hidden assumptions in your answer - are they suitable for the situation in the problem?

Figure 8: The geometry task, Lia's solution, and the teacher's (partial) feedback

Lia's responses to the post-questionnaire presented above reflects an increased understanding of feedback overtime, alongside some struggles to discern the teacher's intended line of thought.

Theme 3: *Enhanced interest and motivation to engage with feedback.* For example, in the pre-questionnaire, Lia wrote:

> I usually feel frustrated because I don't understand what I am supposed to do with the feedback. Or when the comments are not clear. It makes me less inclined to look at the feedback [I feel].

In the post-questionnaire, she wrote:

> I really felt involved, like it was my own thing. I was trying to figure out what went wrong and what I could do differently. It was tough, but I actually enjoyed the challenge. Take the feedback I got, for example, which pushed me to try a different approach. I struggled with it for a while because it was really hard. Then, out of nowhere, I thought to drop equations and use a graph instead. After that, everything started to make more sense and flowed better. So maybe I felt frustration, but also satisfaction." [I feel]

Theme 4: Enhanced self-confidence to engage in feedback, including handling critique. For example, in the pre-questionnaire, Lia wrote:

> I'm unsure how to use the feedback and usually don't feel capable of applying the comments to improve [My confidence].

In the post-questionnaire, she wrote:

> I now believe more strongly that I can use feedback to improve. My confidence in myself has grown, and I'm gradually becoming more comfortable even when receiving critique [My confidence].

As evidenced by Lia’s responses to the pre- and post-questionnaires, there is a shift to a sense of being capable of utilizing feedback effectively.

Summary. The summary indicates a shift among students toward recognizing themselves as integral parts of the feedback process. This shift aligns with key aspects of feedback literacy, as outlined by Carless and Boud (2018), including appreciating feedback, making judgments, managing affect, and taking action. These changes were evident across various dimensions such as knowledge, feelings, beliefs, roles, and confidence, as described by Looney et al. (2018), though potential clashes emerged. For example, students expressed sentiments like, "My role is..., but I feel unconfident," "I have motivation, but partial knowledge," and "I know the criteria, but I am unsure how to apply them." The main explanations for this shift from the students’ perspective include explicit discussions between the teacher and students about feedback and the students' role in utilizing it, as well as empowering students to take an active role in using feedback, providing feedback on the teacher’s feedback, and engaging in reflection.

3.4 Synthesis of findings

Across the studies, the findings indicate that engaging in feedback activities led to a significant shift toward feedback literacy among both teachers and students. For teachers, this shift was observed across various dimensions, including knowledge, feelings, beliefs, roles, and confidence, as described by Looney et al. (2018). This shift manifested in teachers providing more opportunities for students to develop their own feedback literacy. Similarly, students also experienced a shift toward feedback literacy through their engagement in feedback activities, which was expressed in their increased appreciation and utilization of feedback. This, in turn, motivated teachers to continue these efforts, suggesting potential for sustainable feedback practices.

However, some challenges were also identified. While both teachers and students showed greater appreciation for feedback and motivation to use it, there was a notable difficulty and lack of confidence in implementation. Specifically, there were hesitations in applying mathematics assessment criteria, both for teachers and students. These hesitations included questions such as whether there is a “preferred” mathematical model (e.g., algebraic, graphical, verbal), whether explanations and mathematical language should always be prioritized, and how to judge creativity. Teachers and students noted that applying criteria in open-ended math tasks felt more subjective compared to traditional tests.

Additionally, there was evidence of challenges in understanding someone else’s thinking, both for teachers and students. This challenge was particularly pronounced in the context of open-ended math tasks and complex solutions, where understanding thinking that differs from one's own proved difficult. These findings highlight the complexities of implementing effective feedback practices, particularly in the context of mathematics education.

4. Conclusion

Our research has provided some valuable evidence regarding the potential for developing feedback literacy among both students and teachers. By applying Looney et al.'s categories, we were able to identify nuanced complexities within the dimensions of feedback literacy, highlighting how these dimensions interact and manifest differently depending on various factors. Importantly, the challenges specific to mathematics emphasize that feedback literacy is deeply connected to the particular discipline in which it is applied, as noted by Carless and Winstone (2023). This reinforces the need for tailored approaches to developing feedback literacy that consider the unique demands and characteristics of each subject area.

To build on these findings, further research is necessary to explore the interconnections between the different categories related to feedback literacy. Understanding how these categories influence and interact with one another could provide deeper insights into how feedback literacy develops and operates in educational contexts. Additionally, there is a need to investigate the implementation of systematic changes in mathematics classroom-feedback practices. Such research could explore how to effectively integrate feedback literacy into everyday teaching practices in a way that is sustainable and impactful. Finally, it is essential to explore feedback literacy across different contexts and on a larger scale. Large-scale applications and studies across various educational settings could provide a more comprehensive understanding of how feedback literacy can be developed and supported in diverse environments, ultimately contributing to more effective teaching and learning practices.

References

Adarkwah, M. A. (2021). The power of assessment feedback in teaching and learning: a narrative review and synthesis of the literature. *SN Social Sciences*, *1*(3), 75.

Ayalon, M., & Wilkie, K. (2020). Developing assessment literacy through approximations of practice: Exploring secondary mathematics pre-service teachers developing criteria for a rich quadratics task. *Teaching and Teacher Education*, *89*.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5e31.

Boud, D., and E. Molloy. 2013. "Rethinking Models of Feedback for Learning: The Challenge of Design." *Assessment & Evaluation in Higher Education*, *38*(6): 698–712.

Brookhart, S. M. (2013). How to create and use rubrics for formative assessment and grading. Ascd.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325.

Carless, D., & Winstone, N. (2020). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education*, 1-14.

Fujita, T., Jones, K., & Miyazaki, M. (2018). Learners' use of domain-specific computer-based feedback to overcome logical circularity in deductive proving in geometry. *ZDM – Mathematics Education*, *50*(4), 699-713.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81-112.

Jonsson, A. (2013). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education*, *14*(1), 63-76.

Ketonen, L., Nieminen, P., & Hähkiöniemi, M. (2020). The development of secondary students' feedback literacy: Peer assessment as an intervention. *The Journal of Educational Research*, *113*(6), 407-417.

Little, T., Dawson, P., Boud, D., & Tai, J. (2024). Can students' feedback literacy be improved? A scoping review of interventions. *Assessment & Evaluation in Higher Education*, *49*(1), 39-52.

Looney, A., Cumming, J., van Der Kleij, F., & Harris, K. (2018). Reconceptualising the role of teachers as assessors: teacher assessment identity. *Assessment in Education: Principles, Policy & Practice*, *25*(5), 442-467.

Mahfoodh, O. H. A. (2017). "I feel disappointed": EFL university students' emotional responses towards teacher written feedback. *Assessing Writing*, *31*, 53-72.

Rakoczy, K., Pinger, P., Hochweber, J., Klieme, E., Schütze, B., & Besser, M. (2019). Formative assessment in mathematics: Mediated by feedback's perceived usefulness and students' self-efficacy. *Learning and Instruction*, *60*, 154-165.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in education: principles, policy & practice*, 5(1), 77-84.

Smit, R., Dober, H., Hess, K., Bachmann, P., & Birri, T. (2023). Supporting primary students' mathematical reasoning practice: the effects of formative feedback and the mediating role of self-efficacy. *Research in Mathematics Education*, 25(3), 277-300.

Teledahl, A. (2017). Mathematics teachers' assessment of accounts of problem solving. In Dooley, T. & Gueudet, G. (Eds.) (2017). *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education* (CERME 10, February 1 – 5, 2017). DCU Institute of Education and ERME.

Schoenfeld, A. H. (2007). What is mathematical proficiency and how can it be assessed? In A. H. Schoenfeld (Ed.), Assessing Mathematical Proficiency (pp.59-73). New York: Cambridge University Press.

Shute, V. J. (2008). Focus on formative feedback. Review of educational research, 78(1), 153-189.

Van der Kleij, F. M. (2019). Comparison of teacher and student perceptions of formative assessment feedback practices and association with individual student characteristics. *Teaching and Teacher Education, 85*, 175-189.

van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2010). Peer assessment as a collaborative learning activity: the role of interpersonal variables and conceptions. *Learning & Instruction, 20*(4), 280e290.

Wiliam, D. (2011). What is assessment for learning? *Studies in educational evaluation, 37*(1), 3-14.

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), The future of assessment: *Shaping teaching and learning* (pp. 53e82). Mahwah, NJ: Lawrence Erlbaum Associates.

Summative Assessment across three countries

Francesca Morselli[1], Paul Drijvers[2], Alice Lemmo[3], Hans-Georg Weigand[4]

[1] DIMA University of Genova, Italy; morselli@dima.unige,it

[2]Utrecht University, the Netherlands; p.drijvers@uu.nl

[3]Università dell'Aquila, Italy; alice.lemmo@univaq.it

[4]University of Würzburg, Germany; weigand@mathematik.uni-wuerzburg.de

Introduction

This article builds upon the themes discussed during the panel held at the conference, which focused on summative assessment. During the panel, three researchers—Drijvers, Lemmo and Weigand—presented information on three countries (Italy, the Netherlands, and Germany, respectively) and offered insights, reflections, and proposals for future directions. In preparation for the panel, a set of questions was shared with conference participants weeks in advance, allowing them to upload their responses to a Padlet platform.

This first section provides a brief summary of the key themes that emerged from the padlet. Researchers from six countries posted their contributions: China, Greece, Mexico, Singapore, Spain, and Turkey. Their comments helped to sketch a first overview of summative assessment practices, their purposes, and emerging trends in each context.

The subsequent sections present the contributions of the three experts.

First theme: how summative assessment is carried out

In most countries, summative assessment predominantly relies on written tests conducted with traditional methods, such as paper and pencil. These assessments are usually administered at the end of the semester or academic year, with final scores calculated as the average of results from multiple tests or assignments throughout the year. We may note that this approach may limit the exploration of alternative assessment methods.

Second theme: the purpose of summative assessment

The primary objectives of summative assessment across the surveyed countries include: ability streaming and entrance to schools and universities.

Summative assessment for ability streaming is used in contexts like Singapore: this practice aims to group students based on their academic abilities. However, research highlights potential drawbacks, such as stigmatization and a narrowing of curricula for students in lower ability streams (Kramer-Dahl & Kwek, 2010).

Summative assessment is used for entrance to schools and universities in countries such as Turkey, where summative evaluation scores play a critical role in determining student progression to higher education levels. Secondary school scores are used to assign students to high schools, while university placements depend on standardized entrance exams.

Third theme: emerging trends and key issues

The survey highlighted the following emerging trends and key issues related to summative assessment.

The first trend concerns equity and access. Several countries face disparities in assessment outcomes due to socioeconomic and geographical inequalities. For instance, students from rural or underprivileged urban areas in China often encounter additional challenges in accessing quality education. In Greece, efforts to create an inclusive public education system have led to a focus on evaluating students with learning difficulties under the general education framework, emphasizing the social dimensions of assessment.

The second trend is related to assessment stress and mental health. High-stakes assessments contribute to student stress and excessive test preparation. In response, countries like China are adopting new curriculum standards that emphasize holistic assessment and performance-based evaluations. Similarly, Singapore is exploring strategies to balance traditional examinations with broader educational goals, gradually reducing their weight while maintaining accountability.

The third trend concerns the use of Artificial Intelligence (AI). Spain has highlighted the growing influence of AI on educational assessments, particularly regarding the quality and integrity of student-submitted tasks. This technological shift raises questions about maintaining fairness and adapting assessment frameworks to the digital age.

The last trend refers to teacher professional development. Effective implementation of summative assessment reforms requires well-trained educators. However, countries such as Mexico report that teachers often lack the necessary training to adopt new methods, leading to the continued use of traditional approaches. Greece has also emphasized the need for professional development programs to support innovative assessment strategies.

Summative Assessment in the Netherlands

Paul Drijvers

The present state

Summative assessment at the national level in secondary education in the Netherlands has the form of final national examinations at the age of 16 years (vocational education), 17 years (pre-higher education), or 18 years old (pre-university education). For all school types, students' final grades are the average of their grade on the final national examination and the grade for school-based examinations, the latter having different possible formats depending on the schools' policy.

The national examinations for some tracks in vocational education are fully digital (assessment *through* technology, Olsher et al., 2023), but are graded manually by the teachers. For other school types, the final national examination is a paper-and-pen test, in which students can use a calculator or, in case of pre-higher and pre-university education, a non-CAS graphing calculator (assessment *with* technology).

Some reflections

The aim of the 2016 curriculum reform for pre-higher and pre-university education to give higher-order mathematical skills, called mathematical thinking activities, a more prominent place in both teaching and assessment has proven to be hard to implement in the final examinations (Drijvers et al., 2019). How to assess such higher-order skills in the traditional assessment format, in combination with their positions in formative assessment, remains a challenge; a challenge that is not limited to the Netherlands but is widely recognized in our community (Olsher et al., 2023).

Further Developments

At present, a curriculum reform process for all educational levels is taking place in the Netherlands. Of course, the role of digital technology in summative assessment is part of the debate. The common opinion seems to be that graphing calculators are no longer up-to-date technology. One option is to switch to examinations through technology, or to examinations with digital tooling on a tablet or laptop in a so-called bootable client lockdown environment. Content wise, the question is whether or not to include CAS facilities, and if yes, to what school types. Once these questions will be answered and the new curricula will be implemented, the next question of how to deal with artificial intelligence as mathematical tools will be begging to be addressed as well.

Summative Assessment in Italy

Alice Lemmo

The present state

The aspects that principally characterize the Italian assessment policy concern teacher responsibility. The Italian national guidelines state that the responsibility for both periodic and final assessment lies with teachers (MIUR, 2012). In addition, the certification of competences at the end of middle (grade 9) and secondary school (grade 13) is also provided by the teachers.

The Italian standardised assessment (INVALSI) involves all grade 2, 5, 8, 10 and 13 Italian students. INVALSI test is in no way part of school assessment, but it is compulsory for participating in the final examination. It has no formal impact on the teaching-learning and assessing process, each teacher decides independently how to analyse, interpret and use the data collected from his/her students in teaching practice. INVALSI test from grade 8 to grade 13 is computer based, in the next two years this will also be the case for grade 2 and 5 (primary school). Italian standardised assessment therefore involves the use of technology (calculators, digital tools) and is administered through technology (computers/tablet), paper and pen are available if the student needs them.

At the end of middle and high school, all students take a final examination that consists in both written and oral assessment for all disciplines, including mathematics. For grade 8 students both written and oral test are designed and administered by teachers. The final mark consists in the performance in the exams, assessed by teachers. Differently, the final mark at the end of grade 13 is composed by both final exam performance and the assessments of the last three years of school career. In particular, 40% concerns the school career, the other 60% is divided into 1/3 for oral discussion and 2/3 for written examination, in turn divided into two tests. The two written exams are statewide, but teachers administer and evaluate them. Only the second test concerns mathematics, but it is administered only for the scientific high schools' students.

Some reflections

The main aspect of the Italian policy concerns the responsibility and autonomy of teachers in the assessment process at each school level. This could be interpreted because teachers' support guides and assists students in their school career in the same manner as the assessment process should do in a formative perspective. Unfortunately, this may reveal as a point of deep weakness: such autonomy leads to several obstacles and critical issues, if it is not supported and trained.

The use of tests and written exams at school is becoming more and more common (Shepard, 2000); in Italy, written exams and tests seem to represent the main tool for class assessment in mathematics (Amado & Morselli, 2023). This tendency could be perceived as a remedy for inconsistencies in the evaluation or grading practices proposed to teachers and/or as a quest for evaluative objectivity (Shepard, 2000). My personal experience as teacher trainer revealed that Italian teachers seem confused formative and summative assessment with subjective and objective practices. This misconception endorses the assessment of procedural at the expense of conceptual understanding (Hiebert & Lefevre, 1986). This has a strong impact on teaching and learning whereby teaching to the test and consequently learning to the test are increasing. This is confirmed by grade 13 INVALSI test results. Students' data is presented divided in five levels: Level 1 represents students that only deal with elementary knowledge and procedural items, while Level 5 denotes students who tackle problem solving and argumentation items. About 50% of grade 13 students is in Level 1 and 2, only 15% of them belong to Level 5 (INVALSI, 2024). This is also confirmed by OECD PISA 2022 results: more than 50% of Italian grade 10 students belong to level 3 and below (OECD, 2023).

Further Developments

Several Italian researchers are engaged in research on formative assessment (e.g. Cusi, Morselli & Sabena, 2017); others on standardised assessment (e.g. Ferretti, Giberti & Lemmo, 2018). Little is yet done on summative assessment, in particular with tests or written exams (Lemmo, 2023). Concerning this last issue, Italian research is exploiting teaching-learning to the test, proposing tests designed upon processes and assessment grids that aim at competencies (Niss & Højgaard, 2019).

This perspective does not impose to drastically change tasks' design but to modify the way of assessing them. In a procedural approach, teachers could choose that the correct answer in a task counts a certain number of points, and he/she could subtract some points for each error or missing information. The feedback students and teachers receive focus on errors and/or missing answers. In a competency approach, each single task of the test contributes for identifying the level for each competency. For example, even with a wrong answer in some tasks, students could demonstrate ability in using algebraic calculations or properties; in other cases, they could show high communication competency and so on. In this latter case, the focus of the assessment is not on error or omission but on the level of the competencies the student employs.

In conclusion, there is a great gap between the development of research in mathematics education and assessment (Niss, 1992). Research needs to overcome this gap remembering that tests and written exams has an impact on the teaching-learning process and regulates it. In the Thirteenth Congress of the European Society for Research in Mathematics Education, into the Assessment in mathematics education working group, "there was consensus on the fact that assessment is a crucial part of the teaching and learning process and not a mere final "appendix" " (Morselli et al., 2023, p. 5).

Summative Assessment in Germany

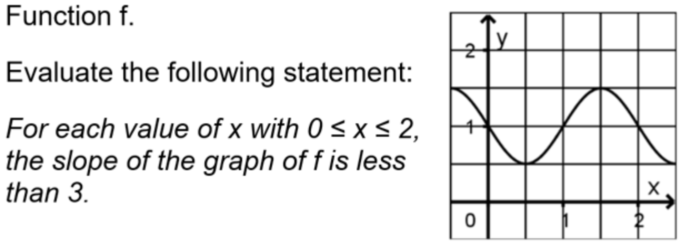
Hans-Georg Weigand

The present state

Concerning the education system, there is no uniform system in Germany. There are 16 states, and each state is responsible for the education system on its own. The consequence is that there is a great variety of school systems and also of final examinations. Although there exist National Standards since 2004, and there is a national "task pool" for exams without technology (Fig. 1), the states can take different tasks from this pool for the first part of the final examination and the second part of the examination is completely different in different states. Concerning digital technologies there are states, which do not allow any digital technology in examinations (except arithmetic calculators), in some other states schools can choose whether they want to do the exam with advanced (usually scientific CAS-calculators) or without technology, and states where CAS-calculators are mandatory for all students.

The federal government tries to control the compliance of the national standards. There are different nationwide annual tests. VERA[1] is a comparative test in grade 3 and 9 (Fig 2 and 3). "IQB-Trends in student achievement"[2] is a test every 5 years in grade 4 and every 3 years in grade 9. It has been a comparative test between the states from 2008-2012, however since 2015 it should "only" give feedback to students, teachers and the policy makers.

Fig. 1 A task from the nationwide "task pool" – final exam



Function f.

Evaluate the following statement:

For each value of x with $0 \leq x \leq 2$, the slope of the graph of f is less than 3.

Fig. 2 VERA - grade 3: How many small boxes does the rectangle have?

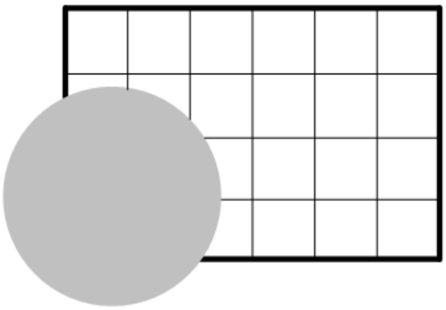
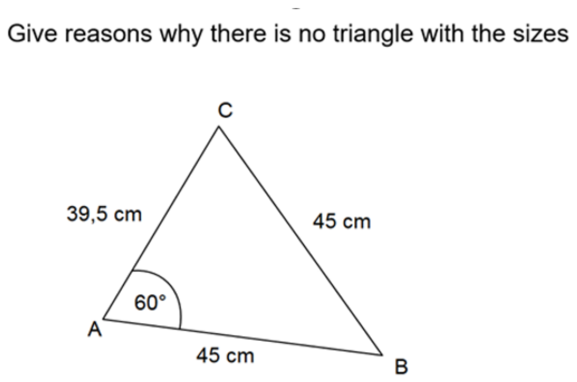


Fig. 3 VERA 8 – grade 9

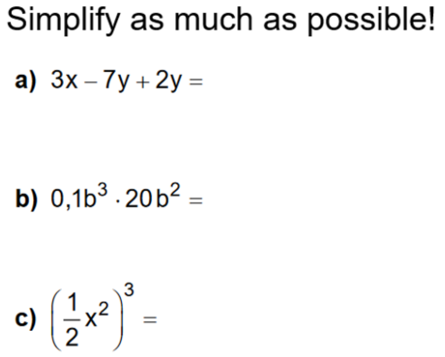
[1] https://www.iqb.hu-berlin.de/vera

[2] https://www.iqb.hu-berlin.de/bt



Give reasons why there is no triangle with the sizes

Fig. 4 Bavarian statewide test - grade 8



Simplify as much as possible!

a) $3x - 7y + 2y =$

b) $0{,}1b^3 \cdot 20b^2 =$

c) $\left(\frac{1}{2}x^2\right)^3 =$

German’s reaction to the (bad) results of TIMMS, PISA, VERA, …

There are – again – quite different reactions to the (bad) results of the international and national tests. E.g., Hamburg or Berlin assigned a scientific working group (2018, 2020). There have been many suggestions for changes concerning the goals, contents and classroom work and the states working on the implementation of some of these suggestions. Bavaria introduced statewide annual – quite classical – tests for grade 8 and 10 at the beginning of each school year (Fig. 4). Last year the nationwide professional development programme “QuaMath” for the next 10 years and for teacher of 10.000 schools was established and at the moment there is just starting a nationwide programme for low achieving students. Sometimes, reactions from the policy makers are done without any empirical basis. In Baden-Württemberg, since 2001 graphing calculators were allowed in examinations, however in 2014, there was a ministry ban and since then no advanced technology is allowed in final examinations. This is an example how arbitrary decisions concerning final examinations are taken.

Further Developments

There are some very important questions and emerging areas for further development concerning summative assessment that are discussed intensively (of course, not only) in Germany. Summative assessment …

- … has a crucial influence on classroom teaching (e.g., how digital media are used).

- … cannot be a 1:1 copy of classroom work. There are some goals which are not – and cannot be – addressed in summative assessment, e.g., modeling or inquiry-based learning.
- … must keep the balance between procedural and conceptual understanding (CU).
- … but must be seen in relation to the development of adequate tasks for CU in the classroom.
- … requires follow-up research on the reasons for good/poor performance.
- … needs strategies to react – in the short and long term – to assessment results.
- … needs to think about alternatives, e.g., oral exams, project work or portfolios.
- … has to be open for up-coming developments like virtual reality, augmented reality or artificial intelligence.

References

Amado, N., & Morselli, F. (2023) Teachers' beliefs about assessment: a study in Italy and Portugal. In In Drijvers, P., Csapodi, C., Palmér, H., Gosztonyi, K., & Kónya, E. (Eds.). (2023). *Proceedings of the Thirteenth Congress of the European Society for Research in Mathematics Education (CERME13).* Alfréd Rényi Institute of Mathematics and ERME. https://hal.science/hal-04413032v1

Cusi, A., Morselli, F. & Sabena, C. Promoting formative assessment in a connected classroom environment: design and implementation of digital resources. *ZDM Mathematics Education* 49, 755–767 (2017). https://doi.org/10.1007/s11858-017-0878-0

Drijvers, P., Kodde-Buitenhuis, H., & Doorman, M. (2019). Assessing mathematical thinking as part of curriculum reform in the Netherlands. *Educational Studies in Mathematics, 102*(3), 435–456. https://doi.org/10.1007/s10649-019-09905-7

Ferretti, F., Giberti, C., & Lemmo, A. (2018). The Didactic Contract to interpret some statistical evidence in mathematics standardized assessment tests. *EURASIA Journal of Mathematics, Science and Technology Education*, *14*(7), 2895-2906. https://doi.org/10.29333/ejmste/90988

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 1–28). Routledge https://doi.org/10.4324/9780203063538

Kramer-Dahl, A., & Kwek, D. (2010). 'Reading' the home and reading in school: Framing deficit constructions as learning difficulties in Singapore English classrooms. In C. Wyatt-Smith, J. Elkins, & S. Gunn (Eds.), Multiple perspectives on difficulties in learning literacy and numeracy (pp. 159–178). Dordrecht: Springer.

INVALSI (2024). *Rapporto Invalsi 2024.* Retrieved from https://invalsi-areaprove.cineca.it/docs/2024/Rilevazioni_Nazionali/Rapporto/Rapporto%20Prove%20INVALSI%202024.pdf

Lemmo, A. (2023). Mathematical written tests as formative assessment practice. In In Drijvers, P., Csapodi, C., Palmér, H., Gosztonyi, K., & Kónya, E. (Eds.). (2023). *Proceedings of the Thirteenth*

Congress of the European Society for Research in Mathematics Education (CERME13). Alfréd Rényi Institute of Mathematics and ERME. https://hal.science/hal-04413550

MIUR (2012). *National guidelines for the curriculum for pre-school and first cycle of education*. [National guidelines for the curriculum for grades K-8]. Annali della Pubblica Istruzione, Numero Speciale. Le Monnier.

Morselli, F., Drüke-Noe, C., Giberti, C., Kaplan-Can, G. & Rämö, J. (2023). An introduction to TWG21: Assessment in mathematics education. *Proceedings of the Thirteenth Congress of the European Society for Research in Mathematics Education (CERME13).* Alfréd Rényi Institute of Mathematics and ERME. https://hal.science/hal- 04412990v1

Niss, M. (Ed.). (1992). *Investigations into assessment in mathematics education: An ICMI study* (Vol. 2). Springer Dordrecht. https://doi.org/10.1007/978-94-017-1974-2

Niss, M., & Højgaard, T. (2019). Mathematical competencies revisited. *Educational Studies in Mathematics*, 102, 9-28. https://doi.org/10.1007/s10649-019-09903-9

OECD (2023), PISA 2022 Results (Volume I): *The State of Learning and Equity in Education*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/53f23881-en.

Olsher, S., Chazan, D., Drijvers, P., Sangwin, C., Yerushalmy, M. (2023). Digital assessment and the "machine". In B. Pepin, G. Gueudet, & J. Choppin (Eds.), *Handbook of Digital Resources in Mathematics Education*. Springer. https://doi.org/10.1007/978-3-030-95060-6_44-1

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4-14. https://doi.org/10.3102/0013189X029007004

Student agency in mathematics assessment: passive targets or active agents?

Juuso Henrik Nieminen[1, 2]

[1]The University of Hong Kong, Faculty of Education, Hong Kong SAR; juuso@hku.hk

[2]Ontario Tech University, Mitch and Leslie Frazer Faculty of Education, Oshawa, Canada

There is no shortage of examples of innovative practices in mathematics assessment research. However, surprisingly little scholarly attention has been given to the presumed role students play in assessment. Students have traditionally been seen as targets of mathematics assessment, deriving from a conceptualisation of assessment as measurement. On the other hand, assessment policies have increasingly emphasised the values of 'Assessment for Learning' that portray students as active agents. This paper proposes a research agenda for understanding student agency in mathematics assessment amid these contradictory trends. By taking a sociocultural and -political approach, I discuss how students largely remain as the 'objects' of assessment, as designed and implemented by others, for the purposes of others. I provide potential ways further if students are to be seen as meaningful agents in mathematics assessment.

Keywords: Mathematics assessment, classroom assessment, student agency, sociology of assessment.

Setting the scene: discursive tensions in how students are positioned in assessment

Mathematics assessment research has vastly unpacked the power of classroom assessment on students' learning of mathematics. There is no shortage of exciting, innovative approaches to assessment in published research literature, particularly when it comes to digital technologies in assessment. A glance at the FAME and CERME TWG21 conference contributions from previous years confirms this: these proceedings include various impressive studies on digital assessment, summative and formative assessment, and standardised assessment, just to name a few key themes. Despite such rich developments in how mathematics assessment should best be designed and implemented, surprisingly, little scholarly attention has been given to the presumed role students play in assessment. Should students be seen as relatively passive targets of assessment who should benefit from mathematics assessment in ways that are determined primarily by educators? Or might it be more beneficial to portray students as active agents who take control of their own learning processes? While these ideas are surely discussed implicitly in a lot of the published work on mathematics assessment, they are commonly overlooked and under-theorised. How could the research community make sense of the student role – students' *agency* – amid the structures of mathematics assessment?

To answer this question, I discuss two global megatrends that surround mathematics assessment, based on Nieminen et al. (2023). These trends find their way into mathematics classrooms at many educational levels, from early to higher education. They shape not only the daily practices of mathematics but the ways in which we think and talk about assessment and student roles within.

First, how mathematics is assessed by teachers at the classroom level is vastly influenced by global testing cultures in which tests are "accepted as foundational practice in education and shape how education is understood in society and used by its stakeholders" (Smith, 2016, p. 10). Indeed, few other school subjects are tested as intensively as mathematics. There is not a similar societal urgency to screen, measure, track and predict students' skills in music, history or geography, even though these subjects are taught widely in schools around the world. Mathematics has particular power in high-stakes testing due to its objective, *testable* nature, which perhaps partly explains why the "measurement paradigm" (Goos, 2020, p. 572) has been so prevalent in mathematics in particular. The rankings produced by international comparison studies, such as the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS), produce national media spectacled around the world (of which the Finnish shock of *losing* to Estonia in mathematics skills in PISA2022 is an excellent example!). Testing and summative examinations are frequent at the classroom level of mathematics, too, which is often attributed to the washback effect of high-stakes testing. However, the "era of testing times" (Marinho et al., 2017, p. 196) is seen in how examinations commonly dominate mathematics assessment in contexts with no high-stakes testing – nor any obligations towards testing – such as higher education (e.g., Iannone & Simpson, 2022) and low-stakes assessment contexts (e.g., Nieminen & Atjonen, 2022).

In the era of testing, students have been seen as "the objects of assessment processes with teachers or external testing bodies controlling the field" (Adie et al., 2018, p. 1). In other words, students rarely have agency over assessment policies and practices, namely, over *how*, *why* and *when* assessment is conducted. This is a particular issue in mathematics due to the high stakes of the discipline (see Nieminen et al., 2023). This was noted by Anne Watson, who, in her excellent book about care in mathematics education, portrayed assessment as an antagonist that *restricts* students' engagement with mathematics:

> ... assessment and accountability systems push schools into a fairly limited range of practices so that many students are trained primarily to pass tests rather than being educated to become competent and confident students and users of mathematics assessment. (Watson, 2021, p. 1)

However, at the same time, recent decades have seen a strong global push towards 'Assessment for Learning' (AfL) policies and practices that emphasise diverse assessment ecosystems stemming from self- and peer-assessment to portfolios to digital innovations to oral assessment (Volante et al., 2024). These 'student-centred practices' are now at the core of mathematics curriculum documents, teacher education programmes, and professional development materials (Goos, 2020). Likewise, such practices are vastly represented in the mainstream of mathematics assessment research.

This push towards understanding mathematics assessment not only as a technology of measurement but as a tool to promote learning seems to promote an *active role* for students. 'Student-centred' assessment aims to activate the students themselves in their own learning processes, enabling them to take more control of their own learning of mathematics. In doing so, students are prepared for the societies of the future, as phrased by Gravemeijer et al. (2017). This idea is present in recent research initiatives on topics such as student agency, activity, engagement, self-regulation, self-efficacy, motivation, control, responsibility, ownership of learning, reflection, higher-order skills, assessment literacy, feedback literacy…

These megatrends may at first seem to contradict each other. From the student point of view, they seem to convey two messages. The era of testing portrays students as compliant objects of tracking and testing: as industrial subjects who should demonstrate their mathematical abilities in identical and comparable ways. On the other hand, AfL positions students as active and empowered 21st century citizens who are unique and self-determined. However, a closer look shows that these two trends intertwine and strengthen each other (for the full story, please refer to Nieminen et al., 2023). Both trends ultimately see students as objects of assessment systems that are defined by experts. For example, many digital formative assessment systems train students to self-regulate their learning of mathematics exactly in ways that are defined by educators, for the purposes defined by educators.

In this way, both megatrends aim to shape students from afar, leaving little room for students' own goals, aspirations, and insights. Students are then seen as objects that can be shaped and moulded through mathematics assessment, which is far from a neutral view (Nieminen & Yang, 2024). In the remainder of this paper, I will propose an alternative by calling for a research agenda on better understanding student agency in and through mathematics assessment.

Towards a research agenda: student agency in mathematics assessment

In my view, if mathematics assessment aims to prepare students for the unknown futures (see e.g., Gravemeijer et al., 2017), there is a need to critically examine how assessment shapes students and their agency as users of mathematics. Given the strong presence of assessment and testing in the day-to-day practices of mathematics education, it is urgent for the mathematics education research communities to unpack the issues of agency in assessment.

Overall, student agency refers to students' capacity to act autonomously and direct in one's life. The idea of agency follows long philosophical, sociological and theological traditions of mapping the relationship between human action and the surrounding social structures. Contemporary conceptualisations of agency tend to differ regarding how much emphasis they place on individuals' own agency amid such structures (see Matusov et al., 2016). Adie and colleagues (2018) contextualised broader discussions around student agency in the context of classroom assessment: 'A focus on student agency in assessment acknowledges students as actors who make choices, and whose actions shape assessment practices in both anticipated and unexpected ways.' (Adie et al., 2018, p. 2) This idea of students shaping the structures of assessment indeed lies at the core of student agency, given that students are not simply 'receivers' of assessment cultures and practices but they actively shape how assessment is conducted and talked about in mathematics (Nieminen & Lahdenperä., 2024; Nieminen & Atjonen, 2023; Nieminen & Tuohilampi, 2020).

Assessment research in mathematics has so far largely focused on individualistic approaches to student agency, examining ideas such as student self-regulation, motivation, self-efficacy, control and choice (see Nieminen et al., 2023). While such views have accumulated an important knowledge base concerning the student viewpoints on assessment, there remains a need to understand the societal, social, cultural, historical, ethical and political underpinnings of the structures of mathematics assessment. Only in this way we can understand the agency of individual students amid the structures that constitute mathematics assessment: assessment policies, institutional practices, social discourses, norms, values, and so forth.

Let me illustrate this point with an anecdote. Consider a rather traditional situation, namely, that a student crams for hours for a mathematics examination during the previous day before the test (you may well imagine this student in your own context, be it primary education in Greece or university mathematics in Australia). The student takes part in the examination and then, regardless of the final test result, forgets most of the covered material immediately. While this may be detrimental to the student's learning of mathematics, life goes on: the teacher moves on to a new topic…

How could the mathematics assessment research community make sense of such (rather common) situations? Imagine further: you are given a massive research grant to study this phenomenon from the viewpoint of student agency. You might take an individualistic viewpoint to agency, since surely the student seems to be using her agency for the maladaptive purpose of cramming. Perhaps you could investigate the student's (apparently lacking) skills in self-regulation or her seemingly low motivation towards learning mathematics for the longer term. However, you might also reach 'beyond the individual' to investigate the structures around student agency. Maybe the immediate structures of assessment promoted such learning behaviours; maybe there was no formative assessment nor opportunities for feedback loops and cycles throughout the learning process? Investigations of social and socio-mathematical norms might reveal how assessment cultures in mathematics education promote cramming. For students, cramming might be a completely normalised way of *doing mathematics*, amplified by popular culture and social media discourses around mathematics. Or, perhaps you would choose to reach even further to investigate assessment policies and testing cultures and how national and international agendas promote certain kinds of learning behaviours at the classroom level. In fact, each of these approaches provides very specific information about students' agency in assessment – and hopefully, if the grant only allows this, all these different viewpoints could come together and supplement each other!

My main point, then, is that the assessment research community should widen its repertoire to understand the social, cultural, political, historical and ethical aspects of mathematics assessment – particularly when it comes to students' agency in assessment. Addressing such diverse viewpoints is by no means a simple task, and indeed Matusov and colleagues (2016) remind us that investigations of agency should draw on complex and multifaceted rather than on one-dimensional approaches. Next, I will discuss two theoretical ideas on how student agency could be conceptualised in mathematics assessment (research). This is, of course, not an exhaustive list but it provides some illustrations of what might be possible.

Authorial agency over the materials of mathematics assessment

It is typically seen as the responsibility of educators, publishers and testing agencies to design the assessment materials and practices of mathematics. For example, there is plenty of academic literature that aims to produce valid and reliable tools for summative and formative assessment of mathematics, be it in the form of scalable tests, digital worksheets, or learning analytics and trajectories (see Nieminen et al., 2023). Based on this approach, assessment materials should be based on expert knowledge. While this is obviously desirable in many cases, at the same time, students are portrayed as non-experts whose voices have little value over how, when and why mathematics is assessed.

Matusov et al's (2016) concept of *authorial agency* is helpful in unpacking how students could be seen as meaningful contributors in mathematics assessment design. When students are provided with authorial agency, they are seen as co-creators of assessment cultures and cultural objects (e.g., test papers, self-assessment forms, assessment instructions, rubrics, syllabus documents…). Thus, rather than indoctrinating students into existing assessment cultures, the idea of authorial agency reminds us that students can be given ownership of such cultures through practices steeped in democratic values. Mathematics classrooms are then seen as learning communities in which all members contribute to the daily practices and materials of teaching, learning and assessment.

There is an increasing knowledge base regarding student partnership in assessment design that might be helpful here (Deeley & Bovill, 2017). Assessment partnership means that students co-design assessment practices together with educators, albeit supported and scaffolded appropriately by educators. As Deeley and Bovill (2017) explain, assessment co-design provides tangible ways to meet the abstract goals of democratic education. This approach turns assessment from mechanical practices into cultural objects in mathematical communities: students may feel ownership of the materials in such communities. In other words, a co-designed rubric is not only a tool for learning, but a cultural object that denotes the values of cultural ownership and democracy. Moreover, when students co-design, for example, formative assessment practices, they might also increase their awareness of how and why assessment is conducted in mathematics (so-called 'assessment literacy').

Since most assessment is at least partly digital now, the lens of authorial agency might also enable educators to foster students' digital agency; their capacity to shape the digital circumstances they live in and to "control and adapt to a digital world" (Passey et al., 2018, p. 426). Students rarely have digital agency in the context of mathematics education. Students are rarely heard when high-stakes testing procedures are digitised; or when new digital technologies are implemented in mathematics classrooms; or when digital platforms are designed to track student learning through learning analytics and learning trajectories. In many ways, students are seen as consumers of technologies in mathematics, particularly when it comes to assessment. In this context, co-designing digital classroom assessment practices – such as digital self-assessment forms, multimodal learning portfolios or annotated rubrics – might result into broader authorial agency when it comes to digital technologies.

At the moment, assessment partnership remains at the margins of mathematics assessment literature. However, Tina Rapke (2016) provides an intriguing example of co-designing a test paper in the context of university mathematics. Students produced test items in groups and, in doing so, felt ownership over the design of this important learning material. What is striking about Rapke's account is that it describes how assessment partnership may be implemented in an institutional context that does not value 'authorial agency' but instead accountability and compliance:

> The idea of involving students in developing the final exam emerged from an attempt to satisfy administrative expectations that students sit a traditional exam (i.e. students would sit an exam individually with only access to a pen or pencil) in my mathematics course, while holding to my philosophy that students and their instructor should play significant roles in the direction of each class and in all assessment practices. (Rapke, 2016, p. 30)

Epistemic agency over mathematical knowledge

Conversation about student agency in mathematics assessment should not disacknowledge the role of mathematics itself in shaping students. Indeed, there is a risk of keeping the discussions around student agency in assessment too general and universal without paying enough attention to the disciplinary knowledge structures of mathematics. From the viewpoint of *epistemic agency*, it could then be asked: how does assessment shape students as mathematical knowers?

The viewpoint of epistemic agency portrays assessment practices as epistemic practices. This means that assessment is conducted in ways that fit the knowledge structures of a given discipline (e.g., mathematical and musical skills are assessed in different ways) and, at the same time, assessment upholds specific epistemologies, or, ways of knowing (Nieminen & Lahdenperä, 2024). Following this approach, the test-driven culture of mathematics assessment could be explained as reflecting the *testable* epistemologies of school mathematics. Moreover, tests and examinations could be seen as upholding certain ways of knowing mathematics as tests commonly rely on individualistic, timed and standardized situations that leave little room for, for example, embodied forms of doing mathematics. In other words, assessment signals students what mathematical knowledge is and is not, and how such knowledge can be assessed.

As epistemic practices, assessment shapes students as epistemic knowers. Here, I rely on the philosophical framework for epistemic justice and injustice by Tanswell and Rittberg (2020) (see also Nieminen & Lahdenperä, 2021). They discuss the ethics of how certain forms of mathematics education shape students as thinkers and doers of mathematics. Broadly, epistemic injustice concerns the injustices of educational institutions that "affect humans specifically in their roles as epistemic agents" (Tanswell & Rittberg, 2020, p. 1200). Epistemic injustice may occur if assessment systematically and institutionally portrays students as 'non-knowers' of mathematics, such as by encouraging cramming cultures as discussed in the anecdote above. Likewise, epistemic injustice may occur if certain ways of knowing are systematically excluded and marginalised through mathematics assessment. For example, what is commonly called ethnomathematics is rarely represented in the assessment ecosystems of assessment, but instead, these ecosystems are built on particular ways of knowing and doing mathematics.

An apt example of epistemic injustice is introduced in Nieminen and Lahdenperä (2021). In this study, one university student criticised examinations as an assessment practice because they promoted only particular ways of being a mathematician, or, an epistemic agent in mathematics:

> A dissertation is a good way of learning, because solving problems and writing about them is what mathematicians actually do. It's not a mathematician's job to sit in a closed exam hall giving answers based on facts and solving procedures learned by heart. So it makes no sense that the education of future mathematicians should prepare you for this kind of activity. (a student in Nieminen & Lahdenperä, 2024, p. 310)

A powerful example of alternative epistemologies in assessment comes from early education in New Zealand (Anthony et al., 2015). This study introduces narrative assessment that captures mathematical learning through *stories*. These stories by children captured elements of mathematical learning that were not otherwise visible and thus promoted students' roles as knowers of mathematics. This example is illuminating since it disrupts the individualistic epistemology of mathematics: the learning stories not only make visible the learning of individuals but supported "educative partnerships with family/whanau" (Anthony et al., 2015, p. 398).

Final words

In this paper, I have argued that the communities of research and practice must pay closer attention to student agency in mathematics assessment. This requires not only psychological but also sociological and ethical explorations of 'agency' that are specific to both mathematics and assessment. I have briefly discussed two potential frameworks for student agency in mathematics assessment – authorial and epistemic agency – but these are only examples of the rich literature on agency waiting to be utilised in the context of mathematics assessment (see Matusov et al., 2016).

Considering student agency in mathematics assessment demands us educators to ask uncomfortable questions from ourselves. How could mathematics assessment *hinder* students' agency in mathematics education? Currently, many assessment practices might simultaneously promote students' learning of mathematics *and* restrict their agency as learners, humans and citizens. To unpack such paradoxical situations, the assessment research community may need to supplement the individualistic and pedagogical paradigms with social, cultural, political, sociological and critical ones, too. For example, well-designed digital platforms for formative assessment may well foster students' learning outcomes in summative tests, but at the same time, they might hinder students' authorial agency by leaving students with no choice other than to partake assessment tasks as designed by educators, test designers or perhaps private companies.

Assessment is commonly cited as the engine of mathematics education reforms (Barnes et al., 2000). This is why I see it imperative to centre the idea of student agency in *assessment* if mathematics education is to prepare students for the societies of the future. That said, student agency is not a normative concept but a risky one since students may exercise their agency in assessment in surprising ways. What kinds of opportunities might emerge if the student agency was placed at the centre of assessment? Exploring this idea might be a risk worth taking.

References

Adie, L. E., Willis, J., & Van der Kleij, F. M. (2018). Diverse perspectives on student agency in classroom assessment. *The Australian Educational Researcher*, *45*(1), 1-12.

Anthony, G., McLachlan, C., & Lim Fock Poh, R. (2015). Narrative assessment: Making mathematics learning visible in early childhood settings. *Mathematics Education Research Journal*, *27*(3), 385-400.

Barnes, M., Clarke, D., & Stephens, M. (2000). Assessment: the engine of systemic curricular reform?. *Journal of Curriculum Studies*, *32*(5), 623-650.

Deeley, S. J., & Bovill, C. (2017). Staff student partnership in assessment: enhancing assessment literacy through democratic practices. *Assessment & Evaluation in Higher Education*, *42*(3), 463-477.

Goos, M. (2020). Mathematics Classroom Assessment. In S. Lerman (Ed.), *Encyclopedia of Mathematics Education* (pp. 572–576). Springer.

Gravemeijer, K., Stephan, M., Julie, C., Lin, F. L., & Ohtani, M. (2017). What mathematics education may prepare students for the society of the future?. *International Journal of Science and Mathematics Education*, *15*, 105-123.

Iannone, P., & Simpson, A. (2022). How we assess mathematics degrees: the summative assessment diet a decade on. *Teaching Mathematics and its Applications: an International Journal of the IMA*, *41*(1), 22-31.

Marinho, P., Leite, C., & Fernandes, P. (2017). Mathematics summative assessment practices in schools at opposite ends of performance rankings in Portugal. *Research in Mathematics Education*, *19*(2), 184-198.

Matusov, E., von Duyke, K., & Kayumova, S. (2016). Mapping concepts of agency in educational contexts. *Integrative Psychological and Behavioral Science*, *50*, 420-446.

Nieminen, J. H., & Yang, L. (2024). Assessment as a matter of being and becoming: Theorising student formation in assessment. *Studies in Higher Education*, *49*(6), 1028-1041.

Nieminen, J. H., & Lahdenperä, J. (2024). Assessment and epistemic (in) justice: how assessment produces knowledge and knowers. *Teaching in Higher Education*, *29*(1), 300-317.

Nieminen, J. H., Bagger, A., Padilla, A., & Tan, P. (2023). Student positioning in mathematics assessment research: A critical review. *Journal for Research in Mathematics Education*, *54*(5), 317-341.

Nieminen, J. H., & Atjonen, P. (2023). The assessment culture of mathematics in Finland: A student perspective. *Research in Mathematics Education*, *25*(2), 243-262.

Nieminen, J. H., & Tuohilampi, L. (2020). 'Finally studying for myself'–examining student agency in summative and formative self-assessment models. *Assessment & Evaluation in Higher Education*, *45*(7), 1031-1045.

Passey, D., Shonfeld, M., Appleby, L., Judge, M., Saito, T., & Smits, A. (2018). Digital agency: Empowering equity in and through education. *Technology, Knowledge and Learning*, *23*, 425-439.

Smith, W. C. (2016). An introduction to the global testing culture. In Smith, W. (Ed.), *The Global Testing Culture: Shaping Educational Policy, Perceptions, and Practice* (pp. 7-24). Oxford: Symposium Books.

Tanswell, F. S., & Rittberg, C. J. (2020). Epistemic injustice in mathematics education. *ZDM*, *52*(6), 1199-1210.

Watson, A. (2021). *Care in mathematics education: Alternative educational spaces and practices*. Springer Nature.

Volante, L., DeLuca, C., Barnes, N., Birenbaum, M., Kimber, M., Koch, M., & Wyatt-Smith, C. (2024). International trends in the implementation of assessment for learning revisited: Implications for policy and practice in a post-COVID world. *Policy Futures in Education*, 14782103241255855.

Full papers

A mathematics teacher's implementation of formative assessment: Overcoming obstacles with adaptive professional development support

Catarina Andersson[1] and Torulf Palm[1]

[1]Umeå University, Sweden; catarina.andersson@umu.se; torulf.palm@umu.se

This paper focuses a mathematics teacher's implementation of formative assessment (FA) when helping students solve mathematics tasks. Such FA practice has great potential, but is non-trivial, and teachers will need substantial support for developing their beliefs and practices. We have studied why an engaged and experienced mathematics teacher who had participated in a comprehensive professional development program made certain changes but not others and how additional support helped her overcome obstacles she experienced. The study exemplifies the significance of first-hand information from teachers' classroom practices together with adapted feedback when providing professional development support for their FA development.

Keywords: Formative assessment, professional development, teacher-student interaction.

Introduction

This paper focuses a mathematics teacher's implementation of formative assessment (FA) when helping students who work individually with mathematics tasks. Such work is frequent in mathematics education in many countries (Hiebert et al., 2003), also in Sweden (Boesen et al., 2014), but providing adequate help to students in this situation is challenging. In this study we follow one experienced primary school teacher who had participated in a comprehensive professional development (PD) program in FA and volunteered to additional individual PD support. We sought to understand why even a committed and experienced mathematics teacher who participated in a comprehensive PD did not make the changes in practice that she desired. We study the changes made, the reasons for making certain changes but not others, and features of the additional PD support. The study contributes to knowledge about crucial features of PD in FA.

Background

Formative assessment, its implementation, and PD support

FA is a classroom practice in which teachers and/or students elicit evidence of students' learning needs through assessment and then adapt teaching or learning to these needs. A large body of research conducted in many different subjects at all educational levels has shown that classroom practices that adhere to the principles of FA can accomplish large gains in student achievement, regardless of whether the teachers or the students were the proactive agents in the FA processes (e.g. Lee et al., 2020). This holds true also for mathematics education (Palm et al., 2017). With this background, not surprisingly, FA has been promoted in many countries.

Despite these promotions, high-quality FA is commonly not enacted in schools. Since such practices include non-trivial aspects of classroom practice, teachers need substantial support for developing both their beliefs about teaching and learning and their practices. Research on PD has identified a number of program features important for attaining desired teacher and student outcomes. Examples are instructional resources, hands-on practice, interactive feedback and discussions, time, and engagement of school leaders and external expertise (e.g. Heitink, 2016). Also, a formative process orientation is pointed out to be a crucial feature of PD programs in FA (Andersson & Palm, 2018).

High-quality formative assessment and its implementation in mathematics education

The present study focuses FA practices where the teacher, as main actor: (1) elicits evidence of student knowledge and skills, (2) interprets the evidence and makes inferences about student learning needs, and (3) gives feedback adapted to these learning needs. The quality of those practices regards the quality of the evidence elicited and the process of using it as information to provide feedback that can support learning. Such FA practices are difficult and complex (Black & Wiliam, 2009), and the use of FA in mathematics accompany specific challenges and changes that rarely come easily (Burkhart & Schoenfeld, 2019). Changes may be giving up intuitive responses such as re-teaching or funneling of students' thinking toward a particular strategy or answer to be replaced by responsive actions which involve taking up and building on students' ideas and thinking (e.g. Jacobs et al., 2022).

Context of the study

The study is part of a three-year intervention project about FA in mathematics. The research group conducted a PD for 25 mathematics teachers at two schools. It was based on previous identified important features of successful PDs (e.g. Heitink, 2016). In the second year, classroom observations revealed that even not all experienced and motivated teachers had implemented all FA practices that was supported in the PD with the quality they aimed for. Additional individual PD support was offered with the intention to target the teachers' specific needs when using FA to help students who work individually with mathematics tasks. The present study focuses on one of these teachers and the following research questions guided the study: Which aspects of the suggested FA practices did the teacher implement and which did she not?; What were the reasons for making certain changes but not others?, and What support helped her to overcome obstacles she experienced?.

Methods

Participants

Elsa (fictious name) volunteered to participate in the study, keen to develop her teaching. She was a committed and experienced mathematics teacher and at the time teaching in Year 6.

Procedure

Elsa audio-recorded her individual help to students during two lessons and sent the recordings to the researchers. From analyzes of the recordings the researchers formulated feedback to Elsa (with examples from the recordings), which was then discussed in a digital meeting between Elsa and one researcher as part of the additional PD. Elsa then used the feedback and attempted to improve her FA practices of her own choice, before a new cycle of co-operation started. The meetings included discussions of Elsa's views of her practice and her beliefs, experiences, difficulties, and successes. The discussions had two purposes: 1) To support her FA implementation; 2) to collect research data.

Data collection

Five teacher-researcher meetings were audio recorded and verbatim transcribed. Other sources of data were the feedback prepared for the meetings, and the recordings from the classroom practice.

Data analysis

To characterize the development of Elsa's FA practices, her classroom recordings were analyzed in relation to the principles and qualitative aspects of FA outlined in the PD. To identify why Elsa made certain changes but not others and to characterize the additional PD support, transcripts from the meetings were analyzed in iterative cycles to identify common themes (see Braun & Clarke, 2006).

Results

Elsa's developed use of suggested FA practices

Before the additional PD support, Elsa struggled to make the students share their thinking and found it difficult to provide feedback that engaged the students in their learning. In her attempts to assess students' needs, several students just answered "I don't know". Elsa tried to push them, using questions such as "What is it that is difficult?". She also used leading questions and asked series of questions with scant room for students to answer. Through her feedback, Elsa tried to encourage the students to take more responsibility but met resistance from the students. During the period of additional PD support, Elsa made progress. For example, she started to use other questions that were easier for the students to answer. She more often insisted on and provided time for students to share their thinking, and occasionally used follow-up questions. Her previous use of feedback pointing out students' successes also got more specific. In addition, from previously taking a leading role in solving the tasks based on her own ideas, she increasingly began to incorporate the students' thinking in her feedback and after giving them hints about how to proceed in their task solving leaving them alone for a while to try to use the feedback.

Reasons for choosing to implement or not implement FA practices

The most salient reasons for not making desired changes had to do with the reactions of the students that Elsa experienced or expected. For example, several students got upset when she insisted that they should share their thinking and Elsa found these students' reactions hard to handle. Moreover, Elsa cared for students' wellbeing and wanted them to enjoy mathematics, which she feared would not be achieved if she demanded too much of them in terms of sharing their thinking or taking responsibility for solving the task. In relation to difficulties of making the students share their thinking Elsa also referred to reasons that had to do with the group constellation. The students did not trust each other and were afraid to reveal shortcomings. Many students with low self-confidence and varying needs made it difficult for Elsa to consistently use the FA practices. Another reason for not using desired FA practices was that she simply forgot, and she referred to the difficulties of abandoning old habits.

Support helping the teacher overcome obstacles

The additional PD provided Elsa with feedback identifying successful uses of the FA practices suggested in the PD, as well as hints about how to develop. Moreover, during the feedback sessions difficulties and reasons (see above) for not using desired FA practices were identified and explored in the dialogue between Elsa and the researcher, and viable ways to implement FA practices were negotiated. Examples of such ways were: introducing changes step by step starting with students who were receptive to changes, providing support and space for students to practice new behaviors, and feedback reinforcing the students' experiences of the benefits of, for example, sharing their thinking.

Conclusion

The research-based PD program provided insufficient support for the use of high-quality FA when helping students individually. The additional PD helped the teacher to make progress regarding both assessment and providing feedback. The study exemplifies the significance of first-hand information of teachers' classroom practices together with adapted feedback for conducting PDs that accomplish large gains in teacher development of FA in the mathematics classroom.

References

Andersson, C., & Palm, T. (2018). Reasons for teachers’ successful development of a formative assessment practice through professional development—a motivation perspective. *Assessment in Education: Principles, Policy & Practice, 25*(6), 576–597. https://doi.org/10.1080/0969594X.2018.1430685

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*, 5–31. https://doi.org/10.1007/s11092-008-9068-5

Boesen, J., Lithner, J., & Palm, T. (2010). The relation between types of assessment tasks and the mathematical reasoning students use. *Educational studies in mathematics, 75*(1), 89–105. https://doi.org/10.1007/s10649-010-9242-9

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Burkhardt, H., & Schoenfeld, A. (2019). Formative assessment in mathematics. In H. L., Andrade, R. E. Bennett, & G. J. Cizek, (Eds.). *Handbook of formative assessment in the disciplines* (pp. 35–67). Routledge. https://doi.org/10.4324/9781315166933

Heitink, M. C., Van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K., & Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educational Research Review, 17,* 50–62. http://dx.doi.org/10.1016/j.edurev.2015.12.002

Hiebert, J. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 video study*. DIaNe Publishing.

Jacobs, V. R., Empson, S. B., Jessup, N. A., Dunning, A., Pynes, D. A., Krause, G., & Franke, T. M. (2022). Profiles of teachers’ expertise in professional noticing of children’s mathematical thinking. *Journal of Mathematics Teacher Education,* 1–30. https://doi.org/10.1007/s10857-022-09558-z

Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The effectiveness and features of formative assessment in US K-12 education: A systematic review. *Applied Measurement in Education, 33*(2), 124-140. https://doi.org/10.1080/08957347.2020.1732383

Palm, T., Andersson, C., Boström, E. & Vingsle, L. (2017). A review of the impact of formative assessment on student achievement in mathematics. *Nordic Studies in Mathematics Education, 22*(3), 25–50.

Enhancing Teacher Training in Mathematics Education: A Model for a Semiotic Approach to Feedback and Interpretative Knowledge

Miglena Asenova[1], Agnese Del Zozzo[2] and Marzia Garzetti[3]

[1]Free University of Bolzano, Faculty of Education, Bolzano, Italy; miglena.asenova@unibz.it

[2]University of Trento, Department of Mathematics, Trento, Italy; agnese.delzozzo@unitn.it

[3]University of Genoa, Department of Mathematics, Genoa, Italy; garzetti@dima.unige.it

This research introduces a comprehensive model for a teacher training course centered on Semiotic Interpretative Knowledge (SIK) in mathematics education. Highlighting the critical need for specialized training, the course is designed to refine teachers' abilities to interpret student's responses through a semiotic lens, especially when conceptual knowledge remains hidden behind difficulties related to patterns of sign use. It focuses on equipping educators with advanced semiotic interpretation skills, thereby enhancing their capability to offer deeper, more meaningful, and effective feedback in mathematics classrooms. The model not only delineates the key features of the designed course but also lays a foundation for future investigations into its feasibility and impact.

Keywords: Teacher training, feedback, semiotic functions, Semiotic Interpretative Knowledge.

Research problem and rationale of the paper

In mathematics education research, the types of feedback that teachers spontaneously provide in their mathematics classrooms are investigated by several studies (e.g., Galleguillos & Ribeiro, 2019; Santos & Pinto, 2010; Stovner & Klette, 2022). The main aspects of feedback that emerge from such research concern conceptual, strategical, or procedural features, while the semiotic aspects related to sign use and production are never explicitly considered. As research shows (e.g., D'Amore & Fandiño Pinilla, 2007; Iori, 2018; Santi, 2011), interpreting student's reasoning requires a strong semiotic competence on patterns of sign use and production. Asenova et al. (2023a) define the notion of Semiotic Interpretative Knowledge (SIK) as "the knowledge needed by teachers in order to interpret students' answers (…), and to give appropriate feedback to them, when conceptual knowledge is hindered, and thus remains hidden behind difficulties related to patterns of sign use and production" (p. 11). In Asenova et al. (2023b) an investigation carried out with 180 Italian prospective primary school teachers shows that prospective teachers spontaneously use a wider web of semiotic resources when they are asked to provide feedback to students, rather than when they are asked to interpret the student's solutions for themselves. Prospective teachers seem to implicitly assume that feedback effectiveness grows with increasing use of semiotic resources, but the above-mentioned study shows also that they often fail in this because of a lack of awareness on the semiotic transformations involved in their feedback. The necessity to consider SIK as a kind of mathematical knowledge for teaching (Ball et al., 2008) and as an explicit content area in teacher training, specifically in reference to feedback effectiveness, goes with a lack of research on how to implement teacher training focused on SIK. This paper presents a model for teacher training courses which address this issue.

Theoretical framework

The semiotic dimension of mathematical knowledge for teaching

Starting from research related to the conceptualization of Mathematical Knowledge for Teaching (MKT) (Ball et al., 2008), Ribeiro and co-authors introduce the notion of Interpretative Knowledge (IK) as the part of the mathematical knowledge "that allows teachers to give sense to pupils' non-standard answers (i.e., adequate answers that differ from the ones teachers would give or expect) or to answers containing errors" (Ribeiro et al., 2016, p. 9).

The semiotic aspects of IK are still little explored, but at the same time research shows that a strong semiotic competence is indispensable for a cognitively meaningful mathematical activity. According to Duval (2017), in mathematics, ostensive references are impossible, as we cannot directly access mathematical objects through our senses. Conceptualization itself, called noesis, is identified in mathematics with a complex coordination of several semiotic systems (called semiosis), rooted in semiotic transformations within the same semiotic system (treatments) and semiotic transformations between different semiotic systems (conversions) (Duval, 2017). A semiotic system (or register) is defined by Duval (1995) and Ernest (2006) as composed by: (1) a set of basic signs that only have meaning when set against or in relation to other basic signs (e.g., the meaning of the digits within the decimal number system); (2) a set of rules for the production of signs, starting from basic signs, and for their transformation. According to Duval, D'Amore (2003) identifies conceptualization with the following semiotic functions, specific to mathematics: (1) choice of the distinctive features of a mathematical object and its representation in a semiotic system; (2) treatment in the same semiotic system; (3) conversion between semiotic systems. The management of such semiotic complexity, within the structure of semiotic systems and the processing of semiotic functions, comes up against Duval's famous cognitive paradox (Duval, 2017): on the one hand, the student comes to know the abstract mathematical objects only through the semiotic activity; on the other hand, such a semiotic activity requires the student's conceptual knowledge of the mathematical objects involved in it. According to Duval, a mathematical object, intended as an epistemic object, arises as the invariant behind treatments and conversions and thus requires the interplay of at least two semiotic registers.

For the reasons displayed above, taking into account the intrinsically semiotic nature of mathematical thinking, Asenova et al. (2023a) introduce the theoretical construct of Semiotic Interpretative Knowledge (SIK) as "the knowledge needed by teachers in order to interpret students' answers (be they standard or non-standard), as well as students' behavior, and to give an appropriate feedback to them, when conceptual knowledge is hindered, and thus remains hidden behind difficulties related to patterns of sign use and production, including individual creativity in sign use" (p. 11). SIK is a kind of MKT that is both subject- and pedagogy-related because the control of the semiotic functions is intertwined both with mathematical knowledge (noesis and semiosis are overlapped) and their implementation in the teaching-learning activity driven by the teacher.

The feedback dimension

Feedback is defined by Hattie and Timperley (2007) as "information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (p. 81). Galleguillos and Ribeiro (2019) investigate prospective teachers' ability to use IK in giving feedback: teachers were asked to solve a task in small groups and then provide feedback to chosen solutions given by students to the same task. These authors classify the provided feedback into four categories: (a) Feedback on how to solve the problem; (b) Confusing feedback: When the feedback seems to be correct, but it can be confusing for the student; (c) Counterexample as feedback; (d) Superficial feedback: The content of such feedback was insufficient (too broad or inconsistent) to allow the solver to understand its meaning. In Asenova et al. (2023b) the authors develop the kinds of feedback introduced by Galleguillos and Ribeiro, consistently with the notion of SIK. In particular, starting from answers given by prospective teacher to similar tasks as the ones proposed by Galleguillos and Ribeiro (2019), the authors categorized the collected feedback according to the implementation of the semiotic functions: type (i) - no mention of semiotic functions, which is framed by Galleguillos and Ribeiro's categories; type (ii) - use of semiotic representations confined to the recognition of the distinctive features; type (iii) - use of distinctive features and treatments; type (iv) - use of distinctive features, treatments, and conversions. The semiotic categorization of feedback does not provide levels of effectiveness per se, but it represents a tool able to tune sign use in producing and evaluating feedback, by identifying levels of complexity of semiotic activity. In this sense, it reduces the gap "between what is understood and what is aimed to be understood" (Hattie & Timperley, 2007, p. 82) on feedback in teacher training and it allows to consider SIK within the categories of mathematical knowledge for teaching.

Research questions and aim

A strong SIK is needed by teachers both to interpret student's solutions, especially when conceptual aspects are hindered by difficulties related to patterns of sign use and production, as well as to provide effective feedback (Asenova et al., 2023a, 2023b). Starting from this assumption and building upon the work of Ribeiro et al. (2016), the present study aims to present a design for a teacher training course on SIK. More specifically, the research question addressed here is: *What are the characteristics of a teacher training course which develops SIK in relation to feedback exchange and production?*

The characteristics of the course will be described, emphasizing their link to the theoretical framework introduced. Additionally, it will be discussed how these characteristics correlate with expectations regarding the processes engaged by the participants throughout the course. The development of a strong SIK on the part of the prospective or in-service teachers is a significant step towards enabling them to understand and be effective in the management and support of student's learning and strategies, particularly in relation to the use of semiotic functions. This proposal serves as a theoretical and methodological basis for future research on the feasibility and effectiveness of the training course described.

The model for SIK operationalization in teacher training courses

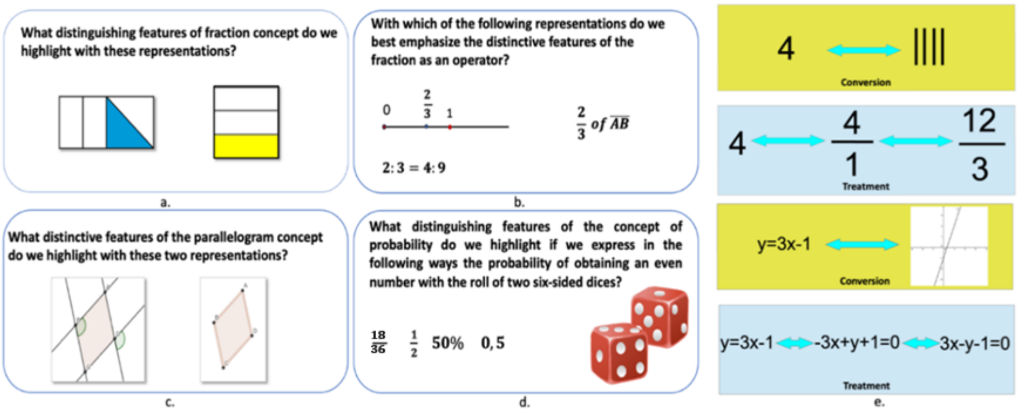
The basic structure of the proposed model is composed of five phases, each associated to an explicit goal (Table 1).

Table 1: Model of a cycle of SIK operationalization in feedback in teacher training courses

	Course phases		**Main goal**
1		Introduction to semiotic functions and their noetic correspondences (step 1) and to semiotic transformations (step 2)	Introduction of tools for the development of SIK
2	Task 1	Task 1: Giving written feedback to student's solutions working in small groups	First implementation of SIK
3	Task 2	Task 2: Written peer-to-peer evaluation to the feedback provided by another group according to criteria related to semiotc functions	Metareflection on SIK implementation in relation to semiotic functions
4		Exchange of feedback and reflection in the groups	Metareflection on SIK evaluation: analysis on one's own work
5		Feedback of the teacher educator during whole class discussion	Institutionalization

In the following the five phases are analyzed into detail and the relationships between the proposed tasks and expectations about the course are unraveled.

Phase 1. During the first phase the participants are introduced to the aspects related to the semiotic functions and their role in the recognition of the mathematical object as invariant behind treatments and conversions. In this phase the definition of conceptualization given by D'Amore (2003) referring to the semiotic functions is strongly used and exemplified. During step 1, focusing on D'Amore's definition of conceptualization, some examples are discussed with the participants recalling their attention to the properties of mathematical objects and the representations of their distinctive features. In Figure 1 four such examples are presented. In the first example (a), choosing the representation on the left highlights the distinctive feature of a fraction as related to equal areas; choosing the representation on the right highlights the distinctive feature of a fraction as related to areas of congruent figures. In the second example (b), only one of the representations ($\frac{2}{3}$ of $\overline{AB}$) highlights the distinctive features of a fraction as an operator (on a magnitude). In the third example (c), choosing the representation on the left brings out the distinctive features of the parallelogram as a quadrilateral with two pairs of parallel sides and, consequently, two pairs of congruent angles; choosing the representation on the right brings out the distinctive feature of the parallelogram as a quadrilateral, but by looking at it as an icon rather than recalling its properties as a bidimensional geometric figure. In the fourth example (d), the initial representation ($\frac{18}{36}$) emphasizes the distinctive features of probability as the ratio of favorable outcomes to total outcomes; the subsequent representation ($\frac{1}{2}$) still portrays probability as a ratio, but does so in a more abstract way: for each of the favorable outcomes, there are two possible outcomes; the third representation (50%) still carries some distinctive feature of probability as a ratio, but only mediated by the meaning of percentage ('per cent' as 'per hundred', from the Latin word 'cento', 'hundred'): there are 50 favorable cases out of 100 possible cases; the fourth representation (0.5) highlights the distinctive feature of probability as a real number in the interval [0;1]. This introduction is especially significant in raising teachers' awareness about the critical attention needed when selecting one representation over another, and the unique characteristics that such representations can either emphasize or conceal.


What distinguishing features of fraction concept do we highlight with these representations?

With which of the following representations do we best emphasize the distinctive features of the fraction as an operator?

What distinctive features of the parallelogram concept do we highlight with these two representations?

What distinguishing features of the concept of probability do we highlight if we express in the following ways the probability of obtaining an even number with the roll of two six-sided dices?

Figure 1: Examples of choice and representation of distinctive features of mathematical objects and of semiotic transformations to be discussed with the participants

Step 2 of phase 1 focuses on semiotic transformation and the participants are introduced to the meaning of treatments and conversions. For this purpose, examples such as those displayed in Figure 1e are presented and discussed with the participants. A conversion between 4 in the decimal arithmetic semiotic system and an iconic representation belonging to the semiotic system of pictograms is accomplished. Then, two treatments within the decimal arithmetic semiotic system are accomplished; in this system there are established rules (how to transform 4 in $\frac{4}{1}$ and $\frac{4}{1}$ in $\frac{12}{3}$) that allow to perform the transformations within the system. In the third row, a conversion between the algebraic semiotic system and the cartesian semiotic system is represented. It allows to evidence the correspondences between the distinctive features of the mathematical object in the semiotic systems: the coefficient '3' corresponds to the slope in the cartesian system; the constant '-1' corresponds to the intercept on the y-axis; each ordered pair (x; y) of numbers that satisfies the equation $y = 3x - 1$ corresponds to a point on the line in the cartesian representation. It is important to point out these aspects because, according to Duval (2017), to support conceptualization, it is not enough to provide representations in different semiotic registers, but it is necessary to make explicit the correspondences between the distinctive features in the different registers. Furthermore, Duval (2017) stresses that conversions are necessary for the conceptualization of the mathematical object, but that the change of semiotic system often makes lose the meaning behind the performed transformations. Teachers often overlook this crucial aspect: they are already familiar with the concept that remains constant across various representations in different systems. However, students frequently miss this invariant and perceive the representations as distinct entities (e.g., students do not recognize a representation of the same object 'line' behind the algebraic and the cartesian representations). But not only conversions lead to a loss of meaning; as D'Amore and Fandiño Pinilla (2007) show, also treatments often lead to a loss of the invariant behind the transformations. Furthermore, it is important to stress that the choice of the distinctive features (properties of the object to be represented) is not independent on the choice of the representation system, as the possibility to represent strictly depends on the semiotic resources provided by the system (D'Amore & Fandiño Pinilla, 2007).

In summary, it is crucial for educators to offer representations in various semiotic systems for two reasons: firstly, identifying an invariant (mathematical object) across different contexts (semiotic systems) requires at least two such systems; and secondly, a single semiotic system is often insufficient to depict all the key features of a mathematical object that students need to understand to fully grasp the concept. The subsequent stages of the model are designed to assist participants in reflecting upon the semiotic perspective introduced within the context of exchanging feedback.

Phase 2. During the second phase of the model, the participants are asked to work in small groups on Task 1 that requires to provide written feedback to student's solution of a task. No specific guidelines are provided on the criteria to be adopted when giving feedback, allowing participants the freedom to use the methods they deem most suitable. It is important to choose tasks that drive the use of semiotic functions, particularly conversions involving symbolic language, natural language, and figural representations. Usually, solutions that contain errors are selected as they more effectively motivate teachers to offer feedback. However, correct yet unconventional solutions that encourage the application of semiotic functions are also suitable. For this purpose, the example proposed by Ribeiro et al. (2016) is particularly suitable (Figure 2), because it involves a nonstandard procedure and the use of only an iconographic register not involving symbolic or verbal ones. Similar problems and their solutions provided by students can be given, involving different registers and strategies.

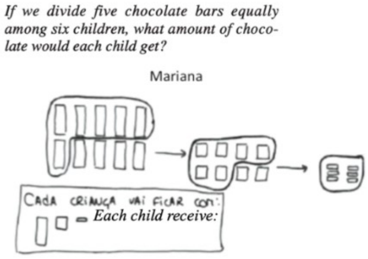

If we divide five chocolate bars equally among six children, what amount of chocolate would each child get?
Mariana
Cada criança vai ficar com:
= Each child receive:

Figure 2: Mariana's solution to the task presented in Ribeiro et al. (2016)

Phase 3. In the third phase of the model, the groups are asked to exchange their feedback with another group and to work on Task 2. This requires them to give a peer-to-peer assessment to the feedback provided by the other group in Task 1. In this case, criteria for the evaluation-feedback are provided. These criteria are the following: (1) Does the choice of distinctive features to represent in the feedback seem appropriate? (2) Does the feedback involve treatments, i.e. semiotic transformations within the same semiotic system? (3) Does the feedback use conversions, i.e. transformations between different semiotic systems? (4) Are the representations chosen in the feedback related to the representation used by the student in the solution? Some examples are provided of how each criterium could be concretely exemplified in relation to the solution provided by the student: for instance, if the teachers want to support the use of a symbolic register in relation to fraction, feedback to Mariana's strategy must go toward this direction. Criteria are provided at this stage to directly foster meta-reflection on using semiotic functions. Additionally, evaluating others' feedback is believed to prompt a detachment from the content, resulting in a more genuine reflection (Grion & Tino, 2018).

Phase 4. In the fourth phase of the model, the groups return their peer-to-peer-feedback to the other group and reflect on the evaluation received from their peers. Providing feedback to another group fosters reflection and offers an opportunity to gain a fresh perspective on the task, enabling to revise and enhance one's own original solving strategies. This is accomplished by reflecting on the evaluation of the appropriateness of the semiotic resources used in providing feedback to the student's solution and on their functionality in supporting the development of the semiotic functions. The feedback provided by others helps to bridge the gap "between what is understood and what is aimed to be understood" (Hattie & Timperley, 2007, p. 82) regarding the appropriateness of one's SIK.

Phase 5. In the fifth and final phase, feedback is provided to the participants by the teacher educator. This occurs in two distinct manners: firstly, through addressing questions that emerge during the activities of the previous phase, and secondly, through the discussion of general issues related to providing suitable feedback to students. Although it is recognized that there is no universally 'correct' feedback and 'ideal solutions' are intentionally not offered, an evaluation on the appropriateness of the implemented semiotic functions is carried out. The objective of this phase is the institutionalization of the new knowledge that has been developed. The focus should here be posed particularly on a comparison of the effectiveness of the chosen semiotic systems and semiotic representations, as well as on the semiotic transformations carried out in providing the feedback and their role in improving one's SIK.

The five phases of the basic model of SIK operationalization in relation to feedback have been presented. The model is conceived as cyclic: the repetition of the model on different problems allows work on semiotic functions in relation to various mathematical objects and varied strategies.

Conclusive remarks

A cohesive model aligned with the theoretical foundations of semiotic functions and IK has been crafted: the characteristics of the course have been presented allowing to work in the direction of defining structured design principles. The course design supports prospective and in-service teachers in consciously referring to semiotic functions while interpreting and providing feedback to students. This approach promotes openness to unconventional or incorrect strategies, highlighting effective conceptual or strategic elements that may be obscured by atypical or incorrect representation usage. The course structure demands focused communicative effort, especially regarding feedback mechanisms and representation choices. It underscores the importance of developing the ability and willingness to understand others' mathematical viewpoints. This comprehension is linked to feedback and interpretations within the SIK framework, recognizing that the selection of a semiotic register reflects an individual's conceptual categories for interpreting reality.

Overall, this contribution lays the groundwork for defining a set of design principles and assessing the effectiveness of courses aimed at developing SIK. It defines the crucial aspects in the foundation of the knowledge needed by the teacher for its development and thus the indicators for its assessment.

References

Asenova, M., Del Zozzo, A., & Santi, G. (2023a). Unfolding Teachers' Interpretative Knowledge into Semiotic Interpretative Knowledge to Understand and Improve Mathematical Learning in an Inclusive Perspective. *Education Sciences, 13*(1), 65. https://doi.org/10.3390/educsci13010065

Asenova, M., Del Zozzo, A., & Santi, G. (2023b). From Interpretative Knowledge to Semiotic Interpretative Knowledge in prospective teachers' feedback to students' solutions. In M. Ayalon, B. Koichu, R. Leikin, L. Rubel, M. Tabach (Eds.), *Proceedings of the 46th of the International Group for the Psychology of Mathematics Education (PME46)* (Vol. 2, pp. 51–58). University of Haifa and PME.

Ball, D.L., Thames, M.H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal for Teacher Education, 59*(5), 389–408. https://doi.org/10.1177/0022487108324554

D’Amore, B. (2003). La complexité de la noétique en mathématiques ou les raisons de la dévolution manquée. *For the Learning of Mathematics, 23*(1), 47–51.

D’Amore, B., & Fandiño Pinilla, M.I. (2007). How the sense of mathematical objects changes when their semiotic representations undergo treatment and conversion. *La matematica e la sua didattica, 21*(1), 87–92.

Di Martino, P., Mellone, M., & Ribeiro, M. (2019). Interpretative Knowledge. In S. Lerman. *Encyclopedia of Mathematics Education* (pp. 1–5). Springer. https://doi.org/10.1007/978-3-319-77487-9_100019-1.

Duval, R. (1995). *Sémiosis et Pensée Humaine: Registres Sémiotiques et Apprentissages Intellectuels*. Peter Lang.

Duval, R. (2017). *Understanding the Mathematical Way of Thinking: The Registers of Semiotic Representations*. Springer.

Ernest, P. (2006). A semiotic perspective of mathematical activity: The case of number. *Educational Studies in Mathematics, 61*(1-2), 67–101. https://doi.org/10.1007/s10649-006-6423-7

Galleguillos, J., & Ribeiro, M. (2019). Prospective mathematics teachers’ interpretative knowledge: Focus on the provided feedback. In U. T. Jankvist, M. Van den Heuvel-Panhuizen, & M. Veldhuis (Eds.), *Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education* (*CERME11*) (pp. 3281–3288). Freudenthal Group & Freudenthal Institute, Utrecht University and ERME.

Grion, V., & Tino, C. (2018). Verso una “valutazione sostenibile” all’università: percezioni di efficacia dei processi di dare e ricevere feedback fra pari. *Lifelong Lifewide Learning, 14*(31), 38–55. https://doi.org/10.19241/lll.v14i31.104

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Iori, M. (2018). Teachers’ awareness of the semio-cognitive dimension of learning mathematics. *Educational Studies in Mathematics, 98*(1), 95–113. https://doi.org/10.1007/s10649-018-9808-5

Ribeiro, C.M., Mellone, M., & Jakobsen, A. (2016). Interpreting students’ non-standard reasoning: Insights for mathematics teacher education. *For the Learning of Mathematics, 36*(2), 8–13.

Santi, G. (2011). Objectification and semiotic function. *Educational Studies in Mathematics, 77*(2-3), 285–311. https://doi.org/10.1007/s10649-010-9296-8

Santos, L., & Pinto, J. (2010). The evolution of feedback practice of a Mathematics teacher. In M.F. Pinto, & T.F. Kawasaki (Eds.), *Proceedings of the 34th Conference of the International Group for the Psychology of Mathematics Education* (PME34) (Vol. 4, pp. 145–152). PME.

Stovner, R.B., & Klette, K. (2022). Teacher feedback on procedural skills, conceptual understanding, and mathematical practices: A video study in lower secondary mathematics classrooms. *Teaching and Teacher Education, 110*(1), 1–12. https://doi.org/10.1016/j.tate.2021.103593

Exploring metrics: elementary mathematics teachers' evaluation of digital geometry assessment activities

Hassan Ayoob[1] and Shai Olsher[2]

University of Haifa, Faculty of Education, Israel; [1]hassan.ayoob@gmail.com; [2]olshers@edu.haifa.ac.il

This study concentrates on metrics that guide teachers in perceiving Digital Formative Assessment (DFA) as suitable for formative assessment for students and aligning with the curriculum. This is part of a larger study that explores the efficiency of a digital formative assessment platform for elementary school geometry. The research utilized open coding to analyze responses from a teacher questionnaire, exploring elementary math teachers' perspectives on DFA. This research was conducted with nine mathematics teachers from diverse Israeli elementary schools. Using a teacher questionnaire and 12 DFA activities, the study identified 11 codes categorized into three categories: Information provided, type of task, and student interaction.

Keywords: Digital formative assessment, elementary math teachers' perspectives, curriculum alignment.

Introduction and theoretical background

Formative assessment is a continuous, multifaceted process integrated into the daily dynamics of teaching and learning. It unfolds through ongoing interactions between teachers and students, wherein teachers adjust their instructional methods and activities based on assessment information (Black & William, 2009). The primary goal is to enhance the learning processes and improve student outcomes. According to Black et al. (2004), the essence of formative assessment lies not only in the assessments themselves, but also in the roles they play in supporting student learning and providing evidence for adapting teaching methods to address specific learning needs. Approaching formative assessment from this functional perspective underscores that its successful implementation hinges on the learning approach adopted and the adept use of knowledge, skills, and strategies by teachers in executing intricate pedagogical processes (Webb & Jones, 2009). For effective formative assessment, teachers' proficiency in regularly collecting and interpreting student learning data is crucial. This complex task extends beyond simple data collection, incorporating timely and constructive feedback, understanding of student learning goals, and tailoring teaching methods accordingly (Black & William, 1998; Heritage, 2007). Central to formative assessment is the role of feedback, which should be continuous, precise, and focused to guide both educators and learners. Feedback is vital for reflective learning, helping students recognize their progress and areas for improvement. Feedback also provides teachers with valuable insights about their teaching methods, by identifying areas where students struggle and understanding students' learning styles. This process helps teachers refine their instructional strategies and make informed decisions to enhance their teaching (Black & William, 1998; Hattie & Timperley, 2007; Heritage, 2007). While many researchers and teachers are aware of the importance of formative assessment for improving the quality of learning and increasing achievement, several factors, such as lack of time for implementation and challenges in adapting tasks, negatively influence formative assessment’s implementation in the classroom (Brown, 2003).

In addition, many researchers agree that teachers' perceptions are one of the main factors impacting implementation in the classroom (Brown, 2003; Black & William, 2009). Assessment, especially formative assessment, is important in the mathematics teaching and learning process, as it helps realize the curriculum's objectives by adapting teaching (Black et al., 2003; Millar, 2016). In recent years, teaching with digital environments in general and geometry in particular has increasingly relied on technology. By offering dynamic forms of mathematical representation of different concepts, dynamic geometry environments (DGEs) provide students with access to mathematical concepts that they have not previously perceived (Leung, 2008; Butler et al., 2010). The suggested DFA activities are designed to effectively support the identification, classification, and analysis of student work methods to the relevant stakeholders (Ayoob & Olsher, 2023). The activities include rich tasks with an infinite number of correct solutions, in the form of the example-eliciting task (EET) (Olsher et al., 2016). These (DFA) tasks feature interactive feedback that highlights the students' exploration and reasoning beyond right or wrong answers. Such design promotes an in-depth interaction with geometric concepts, enabling students to articulate and reflect on their thought processes (Olsher et al., 2016), thereby facilitating a critical examination of their reasoning strategies (Stacey & William, 2013). This type of task involves student-centered assessment that can provide teachers and students with the characteristics of each student's work (Olsher, 2022).

Methodology

The qualitative methodology in this research involves using open coding to analyze responses from an open-ended teacher questionnaire. The focus of this study is to systematically examine elementary mathematics teachers' perspectives regarding computerized activities, specifically digital formative assessment (DFA).

This study is part of a larger research project that explores the efficiency of DFAs for elementary school geometry. Our goal in this report is to describe the metrics according to which elementary mathematics teachers perceive digital formative assessment activities as appropriate for their students and the content of the curriculum. To achieve this goal, we seek to answer the following research questions: What metrics do math teachers use when evaluating the suitability of a digital formative assessment (DFA) to the content and skills detailed in the curriculum?

Research setting

In this study, we analyzed the responses gathered through an open-ended questionnaire. Using an open coding qualitative method, we systematically examined teachers' insights to identify and categorize the metrics they provide. We aimed to address our research question by delineating emerging patterns and themes. This approach enabled us to uncover rich, context-specific information about teachers' perspectives on the effectiveness and alignment of DFA's with curriculum goals and student skills.

Population

Nine mathematics teachers from five different Arabic-speaking elementary schools in northern Israel. All schools teach according to Israel's national curriculum. The participants are a diverse group of teachers with varying educational backgrounds and levels of experience. The majority held master's degrees, with teaching experience ranging from 8 to 25 years, in the 3rd to 6th grades. Collectively, the teachers implemented 69 digital formative assessment activities (and answered the accompanying questionnaires), showing active participation in the research.

Research tools

To examine the metrics according to which elementary mathematics teachers evaluate DFA activities as appropriate for students and the curriculum content, we used a teacher questionnaire and DFA units designed in the STEP-MFA system.

The **12 activities** were designed according to the design principles set out in previous research (Ayoob & Olsher, 2023). The activities align with the textbooks and with the recommended instruction and distribution of the topic to teaching hours according to the curriculum. The tasks were designed in the STEP-MFA environment, and the automatically assessed characteristics of student submissions were designed according to the misconceptions and perceptions in previous literature and the common errors for each subject.

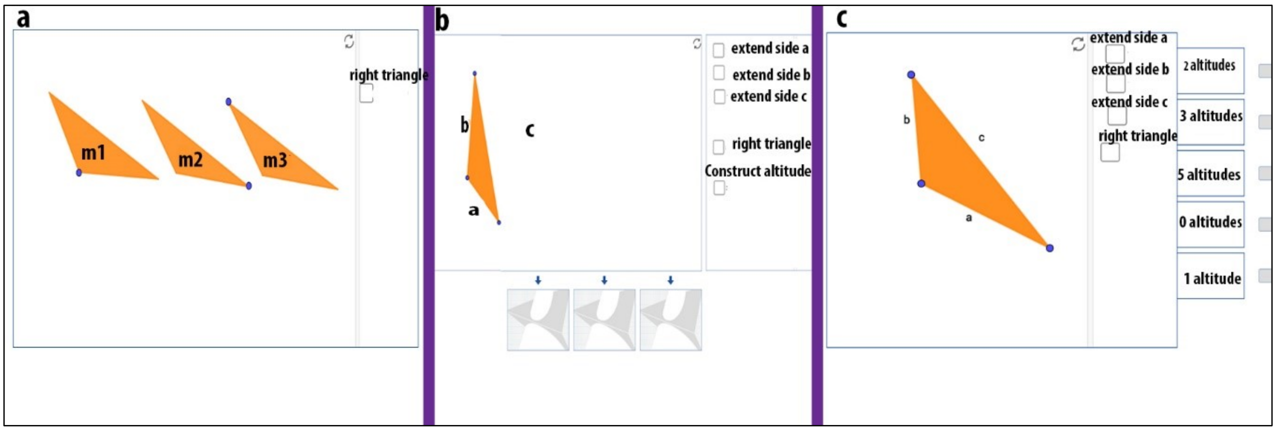


Figure 1: Three tasks of the obtuse triangle altitude activity: (a) construct altitudes from the marked vertex, (b) construct three different examples of obtuse triangles and their altitudes, and (c) determine how many altitudes are outside the triangle?

The following are examples of DFA activity: The mathematical topics studied in this study were the triangle's altitude and polygon area. The activity about the altitudes of an obtuse triangle is shown in Figure 1. The first task (figure 1a) shows three instances of the same obtuse triangle. In each instance, the student is asked to drag the bold vertex to construct an altitude from that vertex and submit it. The triangles are static, and students can drag only the bold vertices. In the second task (Figure 1b), students were asked to construct three different examples of obtuse triangles and then construct one altitude by dragging a point from a vertex. Students can drag the vertices of a triangle to create obtuse triangles of their choices. In the third task (Figure 1c), students are asked to specify how many altitudes can be constructed that are external to the static obtuse triangle and to construct them again by dragging points from the vertices to “stretch” segments. On the left side of each task, students have the option to select and use several tools, such as the option to display the lines extending the sides or a right-angled triangle ruler.

The **questionnaire** consisted of 12 questions, each featuring a dual structure. The first part involves a query with responses based on a linear or multiple-choice format. The subsequent section required the participants to explain their selection. For this study's purpose, we focused on six open-ended questions pertaining to qualitative analysis, concentrating solely on verbal explanations. Question 3 directly addressed indices, inquiring about the metrics used to determine the suitability of a task as a formative assessment activity. Additionally, Question 6 sought an explanation for choosing the tasks deemed most suitable for the students. Question 7 sought an explanation of the rationale behind selecting tasks that were considered unsuitable for students. Question 8 sought an explanation for choosing tasks that provided information about students. Question 9 aimed to understand the reasoning behind choosing tasks that did not offer information about the students. Finally, Question 12 explored the preference for a specific report among the four reports available on the platform.

Research setting

Preparatory stages were undertaken before introducing activities to the classroom for both teachers and students on the STEP platform. Throughout the academic year, our focus was on ensuring that teachers promptly conducted assessments after teaching each sub-topic. The system diligently recorded all the submissions for every student. After the assessment and after any discussions with students about their submissions, the teachers completed the questionnaire. In this questionnaire, they elaborated on their perspective regarding the tasks as formative assessment tasks and explained why they deemed them appropriate.

Following the completion of all activities and the collection of teachers' responses, we initiated a coding process for teachers' answers, categorizing them based on their characteristics. The coding process and questionnaire were validated by the Mathematics Education Research and Innovation Center (MERI) team comprising graduate students, curriculum developers, and researchers in mathematics education.

Data collection and analysis

The data for this study were obtained from the responses of nine teachers to six open-ended questions in the questionnaire, following the completion of each of the 12 research activities.

The data analysis process in this study followed a qualitative approach, primarily utilizing open coding to extract meaningful insights from the responses obtained through a teacher questionnaire. Initially, teachers' answers were transcribed and organized into sheets by question. Open coding involves systematically examining and labeling data segments with descriptive codes, capturing the essence of each unit. Through constant comparison, new data were compared with existing codes to refine the categories and identify patterns and variations. Categories were developed to represent common themes emerging from teachers' insights, progressing from specific codes to more generalized concepts. To ensure the validity of our analysis, a validation process was implemented, cross-checking preliminary findings with the MERI team and validating over 15% of the data. In all cases, disagreement discussions led to consensus and, in some cases, modification of the coding scheme.

Results

In our exploration of teachers' responses to the questionnaire, a comprehensive dataset of 69 statements for each question was collected. The coding process yielded 11 distinct codes.

Subsequently, these codes were aggregated and structured into three overarching categories, as detailed in Table 1. **Information provided**, **type of task**, and **student interaction.** A detailed breakdown of these categories is provided below.

Table 1: Categories and codes for Teachers' responses on the suitability of activities as formative assessment

category	codes
Information provided	Classroom or group information
	Information about the student (Individual)
	Correctness
	Student work methods and misconceptions
Type of task	Curriculum objectives
	Difficulty levels
	Variety of tasks
	Suitability for a diversity of students
Student interaction	Task modification
	Number of submitted examples
	Thought development

Following table 1, and to elucidate the categories and their corresponding sub-categories, we will outline each one followed by its sub-categories. For each sub-category, we'll provide an example of a teacher's verbal response. This approach aims to clarify the rationale behind our categorization and selection process, offering insights into the specific teacher feedback that influenced our decisions.

Information provided

This category captures the depth and variety of information that DFA activities offer teachers about student learning and comprehension. Teachers looked at the information provided by the activity for insights about their class or a group of students. “In my view, if I need to teach the topic again because of students' mistakes, it suggests that the tasks are fitting for assessment activity” (Teacher A, answer to Question 3). Capturing nuanced details specific to individual students. “The activity gives me an in-depth idea of the student's level, especially after completing all the tasks in the activity” (Teacher N, Q 3).

This activity allows teachers to obtain information on the correctness or errors in students' responses. “As a teacher, I will have the knowledge of who has answered correctly and who has not, and I will also be aware of how often certain answers are given. This will enable me to effectively deal with the outcomes” (Teacher B, Q 12). Analysis of the varied methods employed by students in their work methods. “It offers the opportunity to understand all the misconceptions” (Teacher L, Q 8).

Type of task

Teachers assess the relevance of DFA activities based on their alignment with curriculum goals, their challenge level, and the variety they offer. The activity matches and checks the curriculum objectives in every aspect. “The activity effectively checks a student's understanding of whether two segments are perpendicular, not only the segments themselves but also their extensions. It assesses comprehension across various levels: basic understanding, knowledge, identification skills, and application proficiency” (Teacher L, Q 3). Perceived difficulty levels of the assigned tasks. “The activity is suited as it caters to both advanced and less proficient students” (Teacher L, Q 3). Diversity and range of tasks. “The activity includes a variety of tasks designed to cater to and challenge the different skill sets of the students” (Teacher D, Q 3). Appropriateness of tasks for specific student groups. “In this task, students are required to provide three examples. However, from my experience, asking for ten examples is challenging. It is probable that only a few students would be able to present ten examples, but those who do show an ability to generalize” (Teacher D, Q 8).

Student interaction

This category reflects how DFA activities facilitate student engagement through interactive and thoughtful task design. Tasks that enable students to modify the given mathematical context while constructing an example. “The second task allows the student also to build and vary the types of triangles he chooses” (Teacher A, Q 3). Preferences and considerations surrounding the number of prompted examples within a given task. “When a student submits numerous examples, their response becomes clearer and more comprehensible” (Teacher R, Q 8). Tasks for the potential to foster the development of critical thinking skills. “Different examples develop students thinking, so we will have an assessment tool that also enables learning” (Teacher R, Q 6).

Summary and discussion

The categorization of teachers' responses into three primary categories provides an understanding among mathematics teachers of the multifaceted considerations involved in effectively leveraging formative assessment tools, reflecting a deep understanding of the multifaceted considerations required to effectively leverage these tools in education.

Exploring the **information provided** by the assessment task category, teachers acknowledged the importance of designing formative assessments for all students in the classroom or student groups, recognizing the diverse learning needs inherent in different educational stages. The focus on individual student characteristics, especially technological proficiency, underscores commitment to personalized approaches in formative assessment. Additionally, the acknowledgment of addressing and correcting misconceptions highlights the role of formative assessment in fostering accurate understanding and facilitating future teaching and learning.

Within the **type of task** category, teachers articulate their perspectives on the design and structure of formative assessment tasks. Emphasis on task structure indicates an awareness of how instructional design influences student engagement. Awareness of developing students' thinking skills reflects an approach that extends beyond measurement to actively cultivate critical thinking abilities. Furthermore, the acknowledgment of the challenge in balancing difficulty levels underscores the commitment to providing tasks that appropriately challenge students while ensuring accessibility.

The incorporation of a variety of tasks and the consideration of task suitability for specific student populations reveals the aim of creating diverse and inclusive learning experiences.

The **student interaction** category shows that teachers recognize the value of providing flexibility in both the number of examples and the mathematical context that can be altered to fit the interest or level of difficulty preferred by the student. Consideration of the number of examples indicates a thoughtful approach to balancing the assessment’s depth and breadth.

These insights reflect an understanding of formative assessment in line with the principles discussed by Black and William (1998) and Heritage (2007).

In general, these categories contribute valuable insights into the ongoing discourse on best practices in formative assessment, particularly in the context of digital tools and technology integration. These insights reveal a holistic perspective on formative assessment practices. Some of the teachers also showed awareness of the importance of receiving information about the student beyond his mistakes, but also of the work method, which is aligned with the capacities of the STEP platform (Olsher et al., 2016).

In addition to the identified categories, it is important to explore teachers' attitudes towards students' automatic feedback and their integration into formative assessments. Teachers did not refer to this type of feedback in the form of personal reports. This observation is significant considering the established role of timely feedback in formative assessment (Hattie & Timperley, 2007). The absence of a focus on teachers' responses may indicate an area for further exploration and emphasis on teacher training. Additional aspects that did not receive attention in this study included the use of technology to facilitate collaborative work among students, including pairings and peer learning experiences. Understanding how teachers harness digital tools to foster collaborative learning environments and whether they endorse peer-driven feedback processes can offer additional insight into digitally enhanced formative assessment practices. Furthermore, the observation that some teachers showed a tendency to focus primarily on identifying where students went wrong underscores a critical area for development in teacher training. Recognizing and addressing the underlying thought processes behind both correct and incorrect student responses can lead to more nuanced teaching strategies that better support student learning.

References

Ayoob, H., & Olsher, S. (2023, July). Curriculum-aligned digital formative assessment for elementary school geometry. In Thirteenth Congress of the European Society for Research in Mathematics Education (CERME13) (No. 4). Alfréd Rényi Institute of Mathematics; ERME.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.

Black, P., Harrison, C., Marshall, L. B. & William, D. (2004). Working inside the black box: assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 8–21. https://doi.org/10.1177/003172170408600105

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. Educational Assessment, Evaluation and Accountability (formerly: Journal of personnel evaluation in education), 21, 5-31. https://doi-org.ezproxy.haifa.ac.il/10.1007/s11092-008-9068-5

Brown, G. T. (2003, November). *Teachers' instructional conceptions: Assessment's relationship to learning, teaching, curriculum, and teacher efficacy*. At a joint conference of the Australian and New Zealand Associations for Research in Education (AARE/NZARE), Auckland (Vol. 28).

Butler, D., Jackiw, N., Laborde, J. M., Lagrange, J. B., & Yerushalmy, M. (2010). Design for transformative practices. In C. Hoyles & J.-B. Lagrange (Eds.), *Mathematics Education and Technology-Rethinking the Terrain* (pp. 425–437). Springer.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81-112. https://doi-org.ezproxy.haifa.ac.il/10.3102/003465430298487

Heritage, M. (2007). Formative assessment: What do teachers need to know and do?. *Phi Delta Kappan*, *89*(2), 140-145. https://doi-org.ezproxy.haifa.ac.il/10.1177/003172170708900210

Millar, R. (2017). Using assessment materials to stimulate improvements in teaching and learning. In K. Hahl, K. Juuti, J. Lampiselka, A. Uitto, & J. Lavonen (Eds.), *Cognitive and Affective Aspects in Science Education Research* (pp. 31-40). Springer. https://doi.org/10.1007/978-3-319-58685-4_3

Leung, A. (2008). Dragging in a dynamic geometry environment through the lens of variation. *International Journal of Computers for Mathematical Learning*, *13*, 135-157.

Olsher, S., Yerushalmy, M., & Chazan, D. (2016). How might the use of technology in formative assessment support changes in mathematics teaching? *For the Learning of Mathematics*, *36*(3), 11–18. http://www.jstor.org/stable/44382716

Olsher, S. (2022). Te(a)Ching to collaborate: Automatic assessment-based grouping recommendations and implications for teaching. *Proceedings of the 15th International Conference on Technology in Mathematics Teaching (ICTMT 15),* (pp. 214–223). Danish School of Education, Aarhus University. https://doi.org/10.7146/aul.452

Stacey, K., & Wiliam, D. (2013). Technology and assessment in mathematics. In M. A. Clements, A. Bishop, C. Keitel, J. Kilpatrick, F. Leung (Eds.) *Third International Handbook of Mathematics Education* (Vol.27, pp. 721–751). Springer. https://doi.org/10.1007/978-1-4614-4684-2_23

Webb, M., & Jones, J. (2009). Exploring tensions in developing assessment for learning. Assessment in Education: *Principles, Policy & Practice, 16(2)*, 165-184. https://doi.org/10.1080/09695940903075925

Contextuality of Application Tasks in Large-Scale Summative Assessments at Lower Competence Levels for Lower Secondary Education

Sven Basendowski[1] and Gilbert Greefrath[2]

[1]University of Rostock, Germany; sven.basendowski@uni-rostock.de

[2]University of Münster, Germany; greefrath@uni-muenster.de

This article investigates the item pool of measurements of mathematical competencies of students at or below the lowest proficiency level through a qualitative analysis of Germany's 2018 IQB-Trends in Students Achievement. Contrary to conventional views, this article emphasizes a nuanced understanding of the difficulty of application tasks in large-scale assessments, with a focus on contextualization. By introducing a novel task pool designed for students with special educational needs, the study addresses previous limitations in accurately assessing their performance. Qualitative analysis reveals improved authenticity and relevance in the new tasks, particularly in private contexts. The findings highlight the importance of refining authenticity and relevance criteria for application tasks at lower proficiency levels, and provide valuable insights for inclusive education contexts.

Keywords: *Applications, test items, authenticity, relevance, special education.*

Introduction

The results of large-scale summative assessments (LSAs) in mathematics are crucial for the further development of the education system. In recent years, there has been research-based debate about whether and how LSAs can accurately measure the mathematical competencies of students at or below the lowest competency level, as found in assessments such as the Organization for Economic Cooperation and Development's (OECD) Programme for International Student Assessment (PISA). This article presents the results of a qualitative analysis on the contextualization of application tasks in the lower competence range from a German national LSA, the 2018 Trends in Student Achievement, conducted by the Institute for Quality Development in Education (IQB). Contrary to the often-discussed general assessment in literature (Knoche & Lind, 2004), which suggests that application tasks are inherently more difficult than the same mathematical tasks without an application reference, the analysis by Mahler et al. (2020) concludes that the contextualization of application tasks must be considered in a more differentiated way. This finding is crucial for the further development of application tasks in LSAs for an inclusive education system aligning with the principles of the United Nations Convention on the Rights of Persons with Disabilities (UN CRPD). Application-related tasks in LSAs serve an educational goal in an inclusive education system. For this purpose, we analysed characteristic contextual features within the current 2018 IQB-Trends' task pool in Germany.

Relevance of the application reference in LSA

Application reference is of particular importance in many LSAs. International LSAs focus on real-world references when assessing mathematical competencies in adolescence and adulthood. In the

PISA study, mathematical competencies are assessed "in authentic application situations" (Baumert et al., 2001, p. 19). This applies equally to the PISA framework for 2022, which aims to emphasize the relevance of mathematics for students and continues to set tasks in authentic contexts. The application reference is a fundamental feature of tasks in PISA studies and other internationally known LSAs, representing a central educational goal for all students. However, there are indications that application-related tasks are empirically more difficult than comparable inner-mathematical tasks Knoche & Lind, 2004). Contrary to the relevance of application tasks as an educational goal for all students, there is a tendency in Germany that the lower the level of completion of an educational program at lower secondary level, the less often modelling tasks are set (Neubrand, 2007).

Operationalization of the application reference in LSAs

The extent to which the applied tasks vary for different target groups in LSAs has not been analyzed to date. This applies, for example, to PISA tasks, which can be divided into contexts relevant to young people according to different areas of life (OECD, 2023). A distinction is made between contexts that address a personal, professional, social and scientific domain (Reinhold et al., 2019). In addition to the range of relevant contexts to be considered, the LSA frameworks emphasize the authenticity of the test items as a requirement (OECD, 2001). Authenticity refers to an extra-mathematical context that needs to be addressed in the situation using mathematical means. The extra-mathematical context should be authentic and not just constructed for the particular mathematical task. Thus, the use of mathematics in this situation should not be limited to mathematics lessons. Authenticity of tasks is included in various classification schemes and descriptions of modelling tasks (e.g. Maaß, 2010).

Authenticity can be operationalized in terms of different dimensions such as the situation, the question or the information and tools provided (Palm, 2007). When multiple dimensions of a task are authentic, students are more likely to make the necessary real-world considerations for the solution (Palm, 2008). Palm (2008) therefore considers a situation to be authentic if it represents a real task situation and if important aspects of that situation are simulated to an appropriate degree. An authentic task is therefore credible to the learner and realistic in terms of the environment (Palm, 2007). A focus on the authenticity of key dimensions of tasks, such as situation, question and methods, is useful (Turner et al., 2022) and will be used in this article.

Tasks in LSAs at the lower competence levels of the 2018 IQB-Trends

In Germany, there is an institutionalized category of special educational needs for students with learning difficulties and disadvantages (SEN ldd students) as distinct from SEN for students with disabilities (OECD, 2007). The special educational needs of students are "considered to arise primarily from problems in the interaction between the student and the educational context" (OECD, 2007, p. 20) and are manifested by a general and persistent failure to achieve school standards, such as in mathematics. As a result, students with learning difficulties perform in the lower proficiency levels of PISA: 61.5% below proficiency level I, 27.9% at proficiency level I and 10.6% above proficiency level I (Müller et al., 2017). Due to this institutionalization in the German education system, it is possible to specifically examine a group of students that perform within the lower competence range among application tasks.

As part of the 2018 IQB-Trends, a target group-specific task pool for SEN ldd students was developed (Mahler et al., 2020). The IQB-Trends examines a representative sample of all students in grades 4 and 9 without SEN every six years. In 2018, the representative sample also included SEN ldd students and was intended to provide information about the extent to which learners in grade 9 have achieved the competence expectations formulated in the national educational standards (Stanat et al., 2019). Competence level Ib is associated with the minimum requirements to gain the lowest formal secondary school diploma (Kölm & Mahler, 2019).

The new target group-specific development in the 2018 IQB-Trends was deemed necessary, because the task pools and competence structure models used nationally (e.g., Südkamp et al., 2015) and internationally (Müller et al., 2017) had not proven to be empirically suitable for adequately recording the performance of SEN ldd students. In the national studies cited, items from pools for the 4th or 6th grade level were used without taking contextual features into account. As a result, there was no task pool or competence structure model that could measure the performance of SEN ldd students in grade 9 with precision (Müller et al., 2017). The new development was able to empirically demonstrate a satisfactorily valid instrument for comparisons between all learners in secondary I programs except for the highest level (Mahler et al., 2020). In doing so, it focused conceptually on the lower competence areas. Further adaptations focused on processing times and the reduction of language and context barriers in the test material, i.e. by gaining a high degree of authenticity (Mahler et al., 2020). The results suggest that the general valuation that application tasks are empirically more difficult per se needs to be differentiated. The closeness to the real world and the authenticity of the context plays a decisive role here (Mahler et al., 2020). However, no analyses have been conducted to determine how these adaptations were reflected in the final task pool or how they influenced performance. Nevertheless, the results are relevant because modelling competencies are central to the subject of mathematics (KMK, 2004), and this area of competence in LSAs must be ensured for each student in an inclusive education system (UN CRPD).

Research question

The state of the art has highlighted the conceptual, didactic and participatory relevance of application-related items in LSAs in mathematics. The new development of an LSA in the educational trend 2018 provides important indications that, considering specific adaptations to authenticity and the relevance of the application reference, it is possible to develop a valid and reliable task pool that ensures the recording of the performance of young people from different educational pathways in the lower competence area. Given the documented test quality of the newly designed task pool for young people from almost all lower secondary programmes except the highest level, it can be assumed that the adaptations of the newly designed task pool can be applied to any student performing at the lower levels of competence in the IQB-Trends studies. Similarly, there is no consensus in the research on whether, due to their complexity, application-related tasks in mathematical LSAs should be eliminated at the lower proficiency levels for students in lower secondary education.

The aim of this analysis is therefore to characterize the features of context embedded in a task pool that can demonstrably and validly measure performance in application tasks at the lowest competence levels. For this purpose, the items of the newly designed task pool from the 2018 IQB-Trends are compared with those of the previous task pool in terms of the authenticity and relevance of the application reference for each lower competence level. This is of interest to be able to provide indications as to which features of the context are suitable for the improvement of future task pools for LSAs. The research question arises:

To what extent do empirically suitable LSA items with an application reference at the lower competence levels differ from previously used LSA items in terms of authenticity and relevance of the application reference per competence level?

Study Design and Material

A structured qualitative content analysis is recommended for the intended qualitative analysis of the selected didactic criteria and characteristics of context embedded in application tasks (Gläser-Zikuda et al., 2020). For the present study, the standard procedure based on Mayring (2010) was adapted. The deductive category system was scrutinized and modified after the first run of analysis of the material. For this first run through the material, approximately half of the items were randomly selected and prepared for coding with the initial deductive category system in MAXQDA. Coding was carried out independently by two coders.

In this analysis of the 2018 IQB-Trends, the entire task pool, including information on item difficulty and assigned competence level, was provided by the IQB (Basendowski & Greefrath, 2024). However, this data can only be used on the condition that no items are published.

The subject of this structured qualitative content analysis is the set of test items used in the 2018 IQB-Trends (Stanat et al., 2019), which includes both the items and the coding guide. The items of the 2018 IQB-Trends were developed by teachers under the guidance of the IQB. Some items were reused from the 2012 test, while others were supplemented with newly designed items specifically for SEN ldd students. The redesign was prompted by the insufficient measurement accuracy for SEN ldd students identified in previous item pools, as explained above. Before being used in the 2018 IQB-Trends, the items were tested with several hundred SEN ldd students in both general and special schools and then selected. In the 2018 IQB-Trends itself, the 188 newly developed items were used as part of the task pool of a total of 521 items for all students, regardless of the level of the secondary I programme they attended (Mahler, Schipolowski & Weirich, 2019).

In accordance with the given research interest, only items with an application reference were selected. As the redesigned item pool is exclusively at levels Ia, Ib and II, only items at these levels were selected. The total of 128 items selected consists of 88 items from the redesigned pool and 40 items from the existing pool.

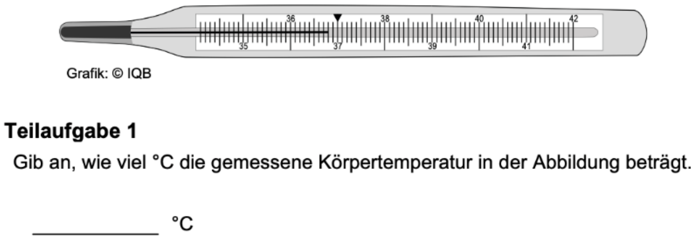
Revision of the category system and final material pass including intercoder reliability check

Gläser-Zikuda et al. (2020) identify testing through intercoder agreement as a common quality criterion for qualitative content analyses and specify a target value of Kappa = 0.70 as a sufficient indicator. The kappa values in the reported study for each supercategory in the second run are between 0.69 and 0.91, determined by MAXQDA. As a result, after the second run there is an outcome that does not necessitate any further revisions of the deductive-inductive category system. All three authenticity categories, i.e. question, situation, and method (Turner et al., 2022), could be coded in "at least simulated authentic" (= credible and realistic), "simulated" (= untrustworthy) and "not assessable". The relevance categories (private, social, professional, scientific) were coded once per item. The final deductive-inductive category system can be found in Basendowski & Greefrath (2024). For an exemplary explanation of the authenticity and relevance criteria, see Table 1 and Figure 1: "Indicate how many degrees C the measured body temperature in the figure is."

Table 1: exemplary coding

Authenticity	question	at least simulated authentic	the problem of measuring a body temperature by reading a clinical thermometer is not implausible
	situation	simulated	reading of a thermometer is not credible without a person being present; in a problem situation, the measured body temperature may have very different consequences
	tools	simulated	the clinical thermometer is virtually available; in a problem situation

Figure 1: Example for previous concept pool items



Grafik: © IQB

Teilaufgabe 1

Gib an, wie viel °C die gemessene Körpertemperatur in der Abbildung beträgt.

__________ °C

(source: https://www.iqb.hu-berlin.de/vera/aufgaben/ma1/)

Results

A comparison between the previous concept and the new concept reveals the following striking differences: The authenticity of the tasks in both pools is generally low. However, there appears to be a trend that the newly designed tasks in the situation category are more consistently simulated authentic than the tasks used in the previous concept. No differences were found in the other authenticity categories. There are major differences in terms of the relevance of the tasks. In both groups, the relevance is essentially limited to scientific contexts, i.e. exemplary questions are described that are not relevant to the students' private, social and future professional lives. However, there are significantly more items among the newly developed tasks that can be classified as relevant in a private context (see Table 2).

Table 2: Number of items in the previous concept and the new concept

		Competence Level Ia		Competence Level Ib		Competence Level II	
		Previous Concept	New Concept SEN ldd	Previous Concept	New Concept SEN ldd	Previous Concept	New Concept SEN ldd
Authenticity: question / situation / tools	not assessable	11 / 0 / 13	36 / 0 / 44	0 / 0 / 0	33 / 0 / 31	25 / 0 / 31	15 / 0 / 14
	simulated	11 / 14 / 8	32 / 41 / 26	2 / 0 / 2	23 / 26 / 22	21 / 31 / 19	6 / 22 / 8
	at least simulated authentic	1 / 7 /1	8 / 35 / 4	0 / 2 / 0	15 / 35 / 18	9 / 22 / 6	6 / 5 / 5

relevance	private	0	18	0	21	0	1
	social	0	2	0	4	0	1
	professional	0	2	0	2	0	0
	scientific	22	54	2	47	50	25

Reading note for lines 2-4: The first number refers to the authenticity category of the question, the second to the authenticity category of the situation and the third to the authenticity category of the tool.

Discussion

The study of Mahler et al. (2020) has revealed that it is indeed possible to develop suitable and valid application tasks at the lowest competence levels. The study indicated that the newly designed task pool allows for a valid assessment of the performance of adolescents at almost all levels of secondary I programs, including the one for SEN ldd students.

It was found that there are clear differences between the concepts. This suggests that the new conceptualisation (Mahler et al., 2020) can provide new insights for the development of LSA items with practical application relevance at lower proficiency levels. In comparison, the new conceptualisation of items for SEN students shows that the novel characteristics 'relevance to everyday life' and 'authenticity of question, situation and tools' of application-related items at lower proficiency levels Ia and Ib were not as prominent in the previous conceptualization of IQB-Trends. It is generally important to enhance the credible problem character of the situation to make it more authentic – even though the new item test pool deviate from the degree of authenticity and relevance intended by the theoretical concept (Mahler et al., 2020).

The absence of authentic contexts on all levels (question, situation, and tools) needs to be discussed in relation to the PISA framework as well (OECD, 2023). There may be a need for improvement. If the demand of the tasks was to increase through authentic and relevant contexts, differences could potentially be identified by comparing the lower competence level with those above. However, in this investigation, no significant differences could be found. This might imply that authentic and relevant contexts can be constructed across all competence levels. Nevertheless, there are indications that authenticity is somewhat less frequent at the lowest competence level, Ia. Therefore, these concerns cannot be entirely dismissed. Other limitations concern items at higher levels of proficiency and more detailed categories.

References

Basendowski, S., & Greefrath, G. (2024). Anwendungsbezug in Mathematik-Large-Scale-Assessments im Bildungsmonitoring für den Sekundarstufe I–Bildungsgang des sonderpädagogischen Schwerpunkts Lernen. Journal für Mathematik-Didaktik. https://doi.org/10.1007/s13138-024-00230-y

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J., & Weiß, M. (Eds.). (2001). PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Leske + Budrich. https://doi.org/10.1007/978-3-322-83412-6

Gläser-Zikuda, M., Hagenauer, G., & Stephan, M. (2020). The Potential of Qualitative Content Analysis for Empirical Educational Research. Forum: Qualitative Social Research, 21(1), 17. https://doi.org/10.17169/fqs-21.1.3443

KMK (Ed.). (2004). Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss. Beschluss vom 4.12.2003. Luchterhand.

Knoche, N., & Lind, D. (2004). Bedingungsanalysen mathematischer Leistung: Leistungen in den anderen Domänen, Interesse, Selbstkonzept und Computernutzung. In M. Neubrand (Ed.), Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland: Vertiefende Analysen im Rahmen von PISA 2000 (pp. 205–226). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-80661-1_11

Kölm, J., & Mahler, N. (2019). Kompetenzstufenbesetzungen im Ländervergleich. In P. Stanat, S. Schipolowski, N. Mahler, S. Weirich, & S. Henschel (Eds.), IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich (pp. 157–168). Waxmann.

Lehmann, R. H., & Hoffmann, E. (Eds.). (2009). BELLA: Berliner Erhebung arbeitsrelevanter Basiskompetenzen von Schülerinnen und Schülern mit Förderbedarf 'Lernen'. Waxmann.

Maaß, K. (2010). Classification Scheme for Modelling Tasks. Journal Für Mathematik-Didaktik, 31(2), 285–311. https://doi.org/10.1007/s13138-010-0010-2

Mahler, N., Kölm, J., & Werner, B. (2020). Entwicklung von Mathematiktestaufgaben für Schüler*innen mit einem sonderpädagogischen Förderbedarf im Lernen – Konzeption und empirische Ergebnisse. In C. Gresch, P. Kuhl, M. Grosche, C. Sälzer, & P. Stanat (Eds.), Schüler*innen mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen (pp. 109–146). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-27608-9_5

Mahler, N., Schipolowski, S., & Weirich, S. (2019). Anlage, Durchführung und Auswertung des IQB-Bildungstrends 2018. In P. Stanat, S. Schipolowski, N. Mahler, S. Weirich, & S. Henschel (Eds.), IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich (pp. 99–124). Waxmann.

Mahler, N., Weirich, S., & Becker, B. (2019). Auswertung, Trendschätzung und Ergebnisdarstellung. In P. Stanat, S. Schipolowski, N. Mahler, S. Weirich, & S. Henschel (Eds.), IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich (pp. 125–130). Waxmann.

Mayring, P. (2010). Qualitative Inhaltsanalyse. In G. Mey & K. Mruck (Eds.), Handbuch Qualitative Forschung in der Psychologie (pp. 601–613). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92052-8_42

Müller, K., Prenzel, M., Sälzer, C., Mang, J., Heine, J.-H., & Gebhardt, M. (2017). Wie schneiden Schülerinnen und Schüler an Sonder- und Förderschulen bei PISA ab? Analysen aus der PISA 2012-Zusatzerhebung zu Jugendlichen mit sonderpädagogischem Förderbedarf. Unterrichtswissenschaft, 2, 194–211. https://doi.org/10.3262/UW1702175

Neubrand, M. (2007). Professionelles Wissen von Mathematik-Lehrerinnen und Lehrern: Konzepte und Ergebnisse aus der PISA- und der COACTIV-Studie und Konsequenzen für die Lehrerausbildung. In: F. Kostrzewa (Ed.), Lehrerbildung im Diskurs (pp. 53–72). gata-Verlag.

OECD (Ed.). (2001). Knowledge and Skills for Life. First Results from the OECD Programme for international Student Assessment. OECD.

OECD. (2007). Students with Disabilities, Learning Difficulties and Disadvantages: Policies, Statistics and Indicators. OECD. https://doi.org/10.1787/9789264027619-en

OECD (2023). PISA 2022 Assessment and Analytical Framework. OECD. https://doi.org/10.1787/dfe0bf9c-en

Palm, T. (2007). Features and Impact of the Authenticity of Applied Mathematical School Tasks. In W. Blum, P. L. Galbraith, H.-W. Henn, & M. Niss (Eds.), Modelling and Applications in Mathematics Education. The 14th ICMI Study (Vol. 10, pp. 201–208). Springer US. https://doi.org/10.1007/978-0-387-29822-1_20

Palm, T. (2008). Impact of authenticity on sense making in word problem solving. Educational Studies in Mathematics, 67(1), 37–58. https://doi.org/10.1007/s10649-007-9083-3

Reinhold, F., Reiss, K., Diedrich, J., Hofer, S. I., & Heinze, A. (2019). Mathematische Kompetenz in PISA 2018 – aktueller Stand und Entwicklung. In K. Reiss, M. Weis, E. Klieme, & O. Köller (Eds.), PISA 2018: Grundbildung im internationalen Vergleich (pp. 187–209). Waxmann. https://doi.org/10.31244/9783830991007

Stanat, P., Schipolowski, S., Mahler, N., Weirich, S., & Henschel, S. (Eds.). (2019). IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich. Waxmann.

Südkamp, A., Pohl, S., Hardt, K., Jordan, A.-K., & Duchhardt, C. (2015). Kompetenzmessung in den Bereichen Lesen und Mathematik bei Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf. In P. Kuhl, P. Stanat, B. Lütje-Klose, C. Gresch, H. A. Pant, & M. Prenzel (Eds.), Inklusion von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen (pp. 243–272). Springer.

Turner, E. E., Bennett, A. B., Granillo, M., Ponnuru, N., Roth Mcduffie, A., Foote, M. Q., Aguirre, J. M., & McVicar, E. (2022). Authenticity of elementary teacher designed and implemented mathematical modeling tasks. Mathematical Thinking and Learning, 26(1), 1–24. https://doi.org/10.1080/10986065.2022.2028225

Assessment through mathematical problem-posing

Rogier Bos and Rebecca Kuijpers

Utrecht University, the Netherlands; r.d.bos@uu.nl

This study concerns problem-posing as a means of assessment in upper-secondary mathematics education. The open character of problem-posing as a task, allows students to show their creativity but makes it difficult to control the focus on the learning goals. Problem-posing can be structured by adding an initial problem to the prompt. We aim to investigate how this form of structuring affects the resulting problems and the extent to which they reveal students' thinking and knowledge with respect to learning goals. In line with previous research on assessment through problem-posing by Kwek and by Mishra and Iyer, we classify the complexity of the problems. Additionally, we analyze whether the problems address the learning goals and are solvable. The main outcome is that structuring the problem-posing prompt is more suitable for assessment since the resulting problems align better with the learning goals and reveal more of the qualities and misunderstandings of the students.

Keywords: Assessment, Mathematics education, Problem-posing.

Introduction

Problem-posing is a teaching approach where students are invited to create or reformulate a problem rather than solve a given one. It has been implemented in some national curriculums and is part of the U.S. Principles and Standards for School Mathematics of the NCTM. Research on problem-posing has been ongoing since at least the ’90s (Silver, 1994; Stoyanova, 1997); see, e.g., the two recent reviews by Baumanns & Rot (2020; 2021). Reasons to teach problem-posing are, among others, that it invites students to analyze situations and that it fosters their creativity (Baumanns & Rot, 2020)

In analogy to the well-known distinction between teaching problem-solving and teaching *through* problem-solving, one can distinguish between assessing problem-posing and assessing *through* problem-posing. The former has been studied extensively (e.g., Silver & Cai, 2005). However, research on assessing through problem-posing is limited to our knowledge (Kwek, 2015; Mishra & Iyer, 2015). If students are taught through problem-posing, as advocated Zhang and Cai (2021), assessing through problem-posing would improve constructive alignment (Biggs & Tang, 2011). The question is: how can we assess a student's learning, knowledge, and skills by considering the problems they pose? The open character of problem-posing tasks seems to prevent teachers from focusing the outcome on the learning goals. Hence the handed-in problems may not allow teachers to assess what they intend to assess. Moreover, the problems students pose may not display the level at which the teacher intended to assess. However, the same open character also allows teachers to assess students’ creativity (Baumanns & Rot, 2021). It may allow students to show what they can do beyond what the teacher might envision, and that could be an attractive property for assessment.

In this study, we investigated how structuring the problem-posing prompt, i.e., providing an initial problem as part of the prompt, might help teachers to nudge the problem-posing in the direction of the desired learning goals and the desired level. It is a trade-off: by structuring the problem the task will be less open, but more focused. Does that hamper creativity? Does it lead to a more informative assessment? By comparing two groups of 10th graders that pose problems, with or without an initial problem, we aim to study how problem-posing can be used for assessment, and whether adding an initial problem improves assessment.

Theoretical background

Stoyanova and Ellerton (1996) differentiate problem-posing situations as free, semi-structured, or structured. The problem-posing starts from a provided real-life or artificial situation. If there are no further restrictions, the situation is called *free*. In a semi-structured situation, the problem should require prescribed mathematical concepts or skills. In a structured situation, students are provided with an initial problem, after which they are invited to pose more problems about the same situation. In the latter case, mathematical concepts and skills are suggested by the initial problem. Baumanns and Rott (2020) often take the two types free and semi-structured together, and so shall we in this article, using the label *unstructured situation*.

We found two papers that explicitly address problem-posing as a means of assessing using semi-structured prompts: Kwek (2015) and Mishra & Iyer (2015). Kwek (2015) introduced rubrics to classify the complexity of the problems that students pose. Table 1 shows a shortened version of these rubrics. Kwek analyzes a set of problems posed by 7th and 9th graders. For grade 9, 78% of the problems are solvable, 67% are of low complexity, 30% are of moderate complexity, and 3% are of high complexity. Students were invited to discuss each other's problems and decide whether these were interesting and challenging. The grade 9 students found 58% of the problems interesting and 50% of the problems challenging. The grade 9 students showed appreciation for problems with strong mathematical content. Kwek concludes that both cognitive factors, like thinking processes and understanding, and affective factors were revealed through classroom problem-posing, making it a suitable assessment activity. However, we believe it could be of additional interest to focus on how the problems reveal students' misunderstanding and to see whether problems cover learning goals.

Mishra and Iyer analyze problems posed as part of a computer science course. Similarly to Kwek, they classify the complexity of the problems based on rubrics: 39% low, 51% medium, and 10% high. 85% of the advanced students who scored high on a classical assessment still produced a problem of medium to low complexity. This indicates that students are not necessarily challenged to perform to their highest ability by a problem-posing assignment. Mishra and Iyer also track which learning goals concerning computational thinking are covered by the problems. They find that some learning goals are better addressed than others, ranging from some goals only covered by 8% of the problems to others by 96%. Missing out on certain learning goals in an assessment can be problematic. In this paper, we study whether providing structured prompts improves such coverage of the learning goals.

In line with these papers, we propose, when assessing through problem-solving, to take into account the solvability of the problem, the complexity of the problem, and the extent to which it covers the learning goals. Solvability is a measure of correctness: if students produce a problem that cannot be solved, this influences the assessment in the same way an incorrect answer influences a traditional assessment. Moreover, it makes sense to take the complexity into account. A problem-posing task has a degree of freedom that could best be compared to the difference between a simple correct answer and an impressive correct answer: complexity is a quantity that allows one to capture this dimension.

As part of the study that we report on in this paper, we compare structured with unstructured problem-posing as a form of assessment. Our research question is: how do these types of situations and prompts for problem-posing contribute to assessment? Is one type more suitable than the other? We restrict ourselves to high-achieving 10th-graders and the subject of probability, but later discuss whether the results might extend beyond these specifics.

Table 1. Complexity of a posed problem; adapted from Kwek (2015)

	Low complexity	**Moderate complexity**	**High complexity**
Description	The problem typically specifies what the solver is to do, which is often to carry out some procedure that can be performed mechanically.	Solving the problem involves more flexible thinking and choice among alternatives. It requires going beyond routine approaches or using multiple steps.	High-complexity problems make heavy demands on solvers, who must engage in more abstract reasoning, planning, analysis, judgment, and creative thought.
Cognitive demand	• Recall or recognize a fact, term, or property • Perform a specified routine of steps • Retrieve information from a graph, table, or figure	• Represent a situation mathematically in more than one way • Justify steps in a solution process • Interpret a visual representation • Extend a pattern • Interpret a simple argument	• Perform a non-routine procedure having multiple steps and multiple decision points • Generalize a pattern • Explain and justify a solution to a problem • Provide a mathematical justification

Method

The study was performed in the context of Mathematics D Online, a Dutch nationwide hybrid (mixed online and onsite) course on advanced mathematics for high-achieving secondary school students. As part of this program, students were invited to hand in answers to a weekly set of tasks. 275 students aged 15 to 16 were enrolled in the course in 2022/2023. We had a sample of 20 students, which we found sufficient, based on previous studies on problem posing with similar data collection (Kwek, 2015; Stoyanova, 1997). The sample was not random, but based on student's positive replies to a request to participate: a convenience sample. We replaced four of the hand-in tasks on probability with problem-posing tasks. Each task consisted of a context and a prompt. For both structured and unstructured tasks, the context was identical, but the prompt differed (see Table 2).

The handed-in problems and accompanying answer models were analyzed and coded for coverage of the learning goals, complexity, and solvability by the second author. The first author performed a second coding. The coverage of learning goals was determined by comparing it to a list of learning goals, representing the material the students were working on. The complexity was coded using an extended version of Table 1. The solvability was determined by carefully examining the problem and the answer model, taking into account that the problem needs to be clear about what needs to be solved and provide enough information to do so, and that it needs to be mathematically correct and consistent. Next, these results were statistically analyzed with suitable tests to allow comparison of structured and unstructured prompts.

Table 2. Examples of problem posing tasks

<table>
<tr><th rowspan="2">Context</th><th colspan="2">Prompt</th></tr>
<tr><th>Structured</th><th>Unstructured</th></tr>
<tr><td>A random variable X has the following distribution:
<table>
<tr><td>X</td><td>0</td><td>1</td><td>2</td></tr>
<tr><td>$P(X = x)$</td><td>$\frac{1}{2} - \frac{p}{2}$</td><td>$p$</td><td>$\frac{1}{2} - \frac{p}{2}$</td></tr>
</table>
with $0 \leq p \leq 1$.</td><td>a. Compute the standard deviation in terms of p.
b. Pose two more problems on this distribution. Also make the answer model.</td><td>Pose three problems on this distribution. Also make the answer model.</td></tr>
<tr><td>A factory produces blue, green, and red soaps. The weight of a soap is normally distributed, with $\mu = 100g$ and $\sigma = 3g$ for blue soaps, $\mu = 120g$ and $\sigma = 4g$ for red soaps and $\mu = 80g$ and $\sigma = 3g$ for green soaps. The volume of a soap is normally distributed, where $\mu = 0{,}2l$ and $\sigma = 0{,}002l$ for blue soaps, $\mu = 0{,}25l$ and $\sigma = 0{,}003l$ for red soaps and $\mu = 0{,}18l$ and $\sigma = 0{,}003l$ for green soaps. The factory sells blue soaps for €1,-, red soaps for €1,50 ,and green soaps for €0,85. The number of soaps sold per day is normally distributed, with $\mu = 40$ and $\sigma = 3$ for blue soaps, $\mu = 35$ and $\sigma = 2{,}5$ for red soaps and $\mu = 40$ and $\sigma = 2{,}5$ for green soaps.</td><td>a. Compute the probability that the volume of a blue soap is less than 0,24l or more than 0,26l.
b. Pose two more problems on this distribution. Also, make the answer model.</td><td>Pose three problems on this distribution. Also make the answer model.</td></tr>
</table>

Results

Nine students handed in 33 problems from structured prompts, and 11 students handed in 53 from unstructured prompts. All problems were coded by the second author. For assessing interrater agreement, 1/3 of the problems were coded by the first author. The codes on solvability and learning goals were identical. Before discussion, there was a Cohen's kappa of 0.66 on the coding of complexity. This was mainly due to confusion about whether a multistep problem should be coded as low or moderate complexity. The raters decided to add the distinction “routine versus non-routine” to the complexity matrix. This decision gave clarity on most differences, leading to a Cohen's kappa of 0.96.

Sample problems

We discuss three problems from our sample here. The first problem was constructed as a response to the first context. The posed problem was: Compute the standard deviation when $p = 1/2$. The problem was coded as having a low complexity because computing the standard deviation is a routine procedure for these students. The problem is solvable and covers a learning goal, namely computing the standard deviation from a probability distribution. Hence the problem posed and the answer model allow us to assess the student's progress for this learning goal, but only on a reproductive level.

The second problem was constructed as a response to the second context. The posed problem was: Compute the probability that the gain of the factory is more than €36, just from the green soap. It was coded as having a moderate complexity because it is not a routine problem. While the computation is a standard one, the solver first has to realize that the answer lies in the number of soaps sold, not the price of the soap. The covered learning goal is: to compute probabilities using the normal distribution, where the average value, standard deviation, and the boundaries are given. The problem was formulated with clarity and is solvable. Hence, it allows us to assess the students on this goal and also conclude that a level of flexibility and creativity was achieved.

The third problem was also formulated in the second context and consists of two parts. Part 1: Which color of soap has a higher probability of being sold less than 39 times per day? Part 2: what are the differences in those probabilities per color? This problem, too, was coded as having a moderate complexity, because the problem, apart from computing probabilities, involves comparing these probabilities. The problem covers the same learning goal as the second problem. The problem is not properly solvable, because of an accidentally too open formulation: the student did not specify how the difference between probabilities should be expressed. While it is unusual to rate the complexity of unsolvable problems (see for example Silver & Cai, 1996), we chose to do so, because with all these problems the intention of the author of the problem was clear, also from their answer model. The formulation of the problem reveals a gap in knowledge about how probabilities should be compared.

Problems covering the learning goals

For each problem posed we analyzed whether it covered at least one learning goal. Of the structured tasks, 97.0% addressed at least one learning goal, whereas for the unstructured versions, this was 90.7%. If a problem did not address a learning goal, then it was in most cases also of low complexity. The learning goals we wanted to be addressed within a context, were addressed in at least some of the problems posed for all learning goals, except one; this exception was due to our fault of not providing a good context for it.

A chi-square test was applied to analyze the differences in coverage of learning goals between structured and unstructured problem-posing exercises. From this, we conclude that there was no significant difference between the number of posed problems that covered the learning goals in structured and unstructured problem-posing exercises ($p = 0.315$).

Complexity of problems posed

To analyze the differences in complexity between structured and unstructured problem-posing exercises, we applied a Mann-Whitney U test to the coded problems. This gave $M - W = 641, p = 0.048$, which means that there is a significant difference in complexity between problems posed as a result of structured and unstructured prompts. The results in Table 3 support that problems posed in response to structured prompts are generally of higher complexity than those posed in response to unstructured prompts.

Problems posed in response to an unstructured prompt tend to be solvable by routine. For example, in the first context a student posed the problems: a. compute the expectation value; b. compute the quadratic deviations; c. compute the variance. The formulation of the structured prompt prevents posing such routine questions because the routine steps are already included as initial problems in the prompt (see Table 2).

Table 3. Relative frequencies of complexity for structured and unstructured exercises

	Complexity			Total
	Low	Moderate	High	
Structured	48,5%	36,4%	15,2%	100%
Unstructured	68,0%	30,0%	4,0%	100%

When students pose complex problems, they do this by combining the context with new elements. For example, one student posed the following problem within the second context: A box contains 20 soaps, of which 6 are blue, 7 are red and 7 are green. What is the probability of grabbing a red soap that weighs more than 125 grams two times in a row (not putting soaps back in the box)? So the student combined discrete and continuous probability.

The solvability of posed problems

We found that 69.7% of the problems posed in response to a structured prompt were solvable, whereas 94.3% of the ones from unstructured prompts were. To analyze the differences in solvability between structured and unstructured problem-posing exercises, we applied a chi-square test. There was a significant difference in solvability between problems posed resulting from structured and unstructured prompts ($p = 0.002$).

The unsolvable problems could be categorized into two categories, namely a category of problems that revealed a misconception or misunderstanding of concepts and a category of problems that were poorly formulated, but otherwise sound. An example of the first category: a student introduces the following discrete probability distribution:

X	0	1	2	3
$H(x = x)$	$1 - h$	h	$h + 1$	$h + 2$

Firstly, we see the notational issues in the first column, which should contain “x” and “$P(X = x)$". This indicates that the student does not understand the role of the random variable X and the variable x. Moreover, adding up the probabilities, one finds $2h + 4$, which should equal 1 (hence $h = -\frac{3}{2}$). However, the student means h not to be determined, and may not realize that the probabilities should add up to 1.

An example of the second category: Compute the revenue of green soaps with a probability of 0,115069670222. This is poorly formulated, and hence unsolvable, since it is not clear what the probability applies to. It could be the probability that a soap is sold to a customer, or the probability that the green soap has a certain weight, etcetera. The answer model revealed a consistent interpretation showing the student’s intention, hence this problem allowed us to assess the student’s progress on both learning content and mathematical problem formulation.

Conclusion and discussion

In conclusion, we state that structured prompts seem more suitable for assessing, for three main reasons. Firstly, structured prompts invite more complex problems, which in turn show more of the students’ capabilities. Secondly, Structured prompts lead to significantly more unsolvable problems than unstructured prompts. This may seem bad, but it is good from an assessment point of view: those problems are usually not fundamentally unsolvable, and how the problems are unsolvable reveals misunderstandings and misconceptions of students. This may be caused by students challenging themselves more with structured prompts. This is in contrast with Mishra and Iyer (2015), who observed that, with semi-structured prompts, students would produce problems below their capabilities, as mentioned in the theoretical background. Thirdly, structured prompts lead to more problems that cover learning goals—though not significantly more. Either way, for both types of prompts the context of the task, combined with the context of the presentation of the task, namely as part of a work on a chapter on statistics and probability, ensured the problems posed revealed students’ progress with respect to the learning goals of the chapter. In most cases, the problems addressed the learning goals we had envisaged, though in some cases this potential was not realized.

Our results were obtained specifically with high-achieving students enrolled in a hybrid national course. Also, the topic was specific: probability. However, we believe the conclusion on the impact of structuring the prompts holds beyond these specifics. Yet, the effect of structuring might wane after students get used to problem-posing, and know what is expected of them. From a new study, we have indications that this may happen within two or three iterations.

Since this study did not involve classroom situations, for future research we would be interested in the impact of exchange between students with respect to the problems they pose, in the line of Kwek’s work (2015). Such discussion might reveal more of students' thinking and invite them to improve their problems.

Acknowledgments

The authors would like to thank Benjamin Rott for suggestions and discussions concerning this study.

References

Baumanns, L., & Rott, B. (2020). Rethinking Problem-Posing Situations: A Review. *Investigations in Mathematics Learning, 13*(2), 59–76. https://doi.org/10.1080/19477503.2020.1841501

Baumanns, L., & Rott, B. (2021). Developing a framework for characterizing problem-posing activities: a review. *Research in Mathematics Education, 24*(1), 28–50. https://doi.org/10.1080/14794802.2021.1897036

Biggs, J & Tang, C. (2011). *Teaching for Quality Learning at University*, McGraw-Hill and Open University Press, Maidenhead.

Kwek, M.L. (2015). Using Problem Posing as a Formative Assessment Tool. In: Singer, F., F. Ellerton, N., Cai, J. (Eds.) *Mathematical Problem Posing*. Research in Mathematics Education. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-6258-3_13

Mishra, S., Iyer, S. (2015). An exploration of problem posing-based activities as an assessment tool and as an instructional strategy. *Research and Practice in Technology Enhanced Learning*, *10*(5). https://doi-org/10.1007/s41039-015-0006-0

Silver, E. A (1994). On mathematical problem posing. *For the Learning of Mathematics, 14*(1), 19–28.

Silver, E. A., & Cai, J. (1996). An Analysis of Arithmetic Problem Posing by Middle School Students. *Journal for Research in Mathematics Education*, *27*(5), 521-539. https://doi.org/10.2307/749846

Silver, E. A., & Cai, J. (2005). Assessing students' mathematical problem posing. *Teaching Children Mathematics*, *12*(3), 129–135. https://doi.org/10.5951/TCM.12.3.0129

Stoyanova, E., & Ellerton, N. F. (1996). A framework for research into students' problem posing in school mathematics. In P. C. Clarkson (Ed.), *Technology in mathematics education* (pp. 518– 525). Melbourne, Mathematics Education Research Group of Australasia

Stoyanova, E. (1997). *Extending and exploring students' problem-solving via problem-posing*. [Doctoral dissertation, Edith Cowen University]. Research Online Institutional Repository. https://ro.ecu.edu.au/theses/885/

Zhang, H., & Cai, J. (2021).Teaching mathematics through problem posing: insights from an analysis of teaching cases. *ZDM Mathematics Education* 53, 961–973. https://doi.org/10.1007/s11858-021-01260-3

Designing with the Teaching for Robust Understanding framework: indicators for the activation and realization of formative assessment strategies

Alessandra Boscolo[1], Francesca Morselli[2], Simone Quartara[3] and Elisabetta Robotti[4]

[1]DIMA University of Genova, Italy; boscolo@dima.unige.it

[2] DIMA University of Genova, Italy; morselli@dima.unige,it

[3]IIS Italo Calvino, Genova, Italy; simone.quartara@calvino.edu.it

[4]DIMA University of Genova, Italy; robotti@dima.unige.it

The study explores the implementation of formative assessment strategies in the context of algebraic thinking and argumentation within a teaching experiment. The Teaching for Robust Understanding framework and theoretical references guide the task design. Specific indicators for two formative assessment strategies are developed, tailored to the learning goals, and examples of their instances (activation and realization by teachers and students) are provided. Future work will extend this analysis to other strategies and assess its applicability in other learning sequences.

Keywords: Formative assessment, TRU framework, community of inquiry

Introduction and background

Our contribution addresses task design and the assessment of algebraic thinking and argumentation as key learning objectives. We rely on Black and Wiliam’s (2009) characterization of formative assessment as a method of teaching where “evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited”. (p. 7). Schildkamp and colleagues (2020) underscore the challenges teachers face in implementing formative assessment effectively within their classrooms. In their literature review, they point out that it is imperative to integrate formative assessment seamlessly into the teaching and learning process, surpassing the mere addition of formative assessment activities. Moreover, teachers should be inclined to share the responsibility of instruction with students, thereby renegotiating the role and authority of the teacher in the teaching and learning process. Teacher prerequisites supporting this shift in perspective and practice encompass pedagogical content knowledge (essential for identifying student difficulties and offering feedback), the ability to articulate and share learning goals with students, and the capacity to facilitate class discussions. Additionally, social factors, such as collaboration with colleagues and cultivating positive relationships with students, play pivotal roles. All these factors underscore the importance of making teachers able to effectively implement formative assessment in their daily classroom practices, recognizing the complexity inherent in the teaching and learning process. We contend that creating an appropriate context for teachers to reflect on their practice, share and compare experiences, and providing them with theoretical tools supporting the design and implementation of teaching sequences, including formative assessment activities as integral components, is paramount.

Our study was conducted within the community of inquiry DIVA (Didattica, Inclusione, Valutazione formativa, Argomentazione – Didactics, Inclusion, Formative Assessment, Argumentation), established at the Mathematics Department of the University of Genoa in February 2023. This community of teacher-researchers has been collaborating to identify theoretical tools for reflection, address specific needs and areas of interest, and design teaching and learning sequences. The initial theoretical tool shared and utilized was Schoenfeld's TRU framework (Teaching for Robust Understanding) (2016) that identifies five dimensions for learning: mathematics, cognitive demand, equal access to content, agency, ownership and identity and formative assessment. The mathematical dimension is at the core of the model and the other dimensions are shaped around it. The dimensions do not contain prescriptive "recipes" for teachers but rather guidelines for creating powerful learning environments, that result in students becoming resourceful thinkers and learners. The dimensions provide an analytical tool for the observation and reflection on one's own teaching practice and can be used for designing and evaluating the effectiveness of the intervention and thinking about the next steps in teaching action.

In this contribution, we present a teaching and learning sequence conceived and implemented within the TRU framework and we study to what extent formative assessment strategies were implemented. Wiliam and Thompson (2007) discuss five key strategies that may help promoting formative assessment in the classroom: FA1) clarifying and sharing learning intentions and criteria for success; FA2) engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding; FA3) providing feedback that moves learners forward; FA4) activating students as instructional resources for one another; FA5) activating students as the owners of their own learning. Not only the teacher, but also the peers and the student himself/herself may act as agents of formative assessment.

Research Design

Our exploratory study is based on a teaching experiment, involving a teacher (the author SQ) who is a member of the DIVA community of inquiry. Despite being an experienced teacher, who already took part in teaching experiments concerning formative assessment (Morselli & Quartara, 2023), this represents his first attempt to design a learning activity through the framework TRU. The activity took place in a grade 9 class (18 students) of an upper secondary school with a scientific orientation.

With reference to the mathematical content dimension, which is at the core of the TRU framework, the activity at issue was aimed at the development of algebraic thinking and of argumentative competence, with a strong focus on the interaction between them. As mentioned earlier, what sets DIVA apart is its approach to design, guided by the theoretical tool TRU, and the practice of sharing explicit reflections with teachers in the community. These reflections are often guided by additional theoretical tools that cater to the various dimensions of TRU and the specific content that their proposals intend to cover. In this case, additional theoretical tools refer to algebraic thinking and argumentation.

Algebraic thinking is explicitly linked to Arcavi's conceptualization of symbol sense (1994). The development of symbol sense involves: understanding how and when to use symbols to represent relationships, generalizations, and proofs; being aware that in some cases it is more convenient to abandon symbols in favor of other approaches; dealing with the dialectic between manipulation and interpretation of symbols; being aware of the possibility of creating symbolic expressions, and being able to create them; being able to select, but also to abandon or change a symbolic representation; being aware of the need to constantly check symbol meanings during problem-solving; being aware of the fact that symbols may play different roles. Concerning the second objective, that is, the development of argumentative competence, we refer to Habermas's characterization of "rational behavior" (Morselli & Boero, 2010), thus identifying three components: epistemic (inherent in the correctness of the argumentative process); teleological (inherent in the problem-solving character of the process, and in the related strategic choices); communicative (related to the comprehensibility and communicative choices of the argumentation). All the specific theoretical references were shared with the teacher before starting the design and implementation.

Once the mathematics dimension was fixed, the design of the activity was structured to take into account the other dimensions of the TRU framework. Due to space constraints, we summarize in Figure 1 the structure of the activity, involving 4 stages, specifying the TRU dimensions motivating the introduction of each stage. We will defer this description to future work. The activity is based on the resolution of an algebraic item selected from the INVALSI national assessment repository GESTINV3 (D14 G10 year 2010). In the first individual stage (10 minutes), students were asked to explore and conjecture around the following open-ended question: "If n is any natural number, what do you get by adding the three numbers 2n+1, 2n+3, 2n+5?". Consequently, they were asked to conjecture the truth value of the following statements: Mario’s ("You always get the triple of one of the three numbers"); Luisa’s ("You always get an odd number"); Giovanni’s ("You always get a multiple of 3"). In the second stage, students were asked to discuss in small homogeneous groups and compare their own conjectures, formulated in the previous stage, with group members, answering the following multiple-choice question: "Who is right? a. All of them, b. Only Mario, c. Only Luisa, d. Only Giovanni.” Students were required to come to a consensus on a solution and produce a written argument in which their solution is accompanied by a justification regarding the truth value of each of Mario, Luisa and Giovanni's statements. In the third stage, involving a whole class discussion, the teacher displayed the students' responses on the whiteboard and asked each group to narrate the solutions and arguments they had previously developed, involving the students from other groups as well to ask, comment, and compare strategies. During the discussion, aspects related to the algebraic correctness of the solutions were examined, along with the formulation of arguments, and the role that algebra had played in the diverse solutions and argumentations. Finally, the students were asked to complete a self-assessment questionnaire, with a Google module, containing questions aimed at monitoring the aspects of the five TRU framework dimensions on which the design focused. All the discussions were video-taped and transcribed. All the students’ productions were collected.

[3] https://www.gestinv.it/Index.aspx

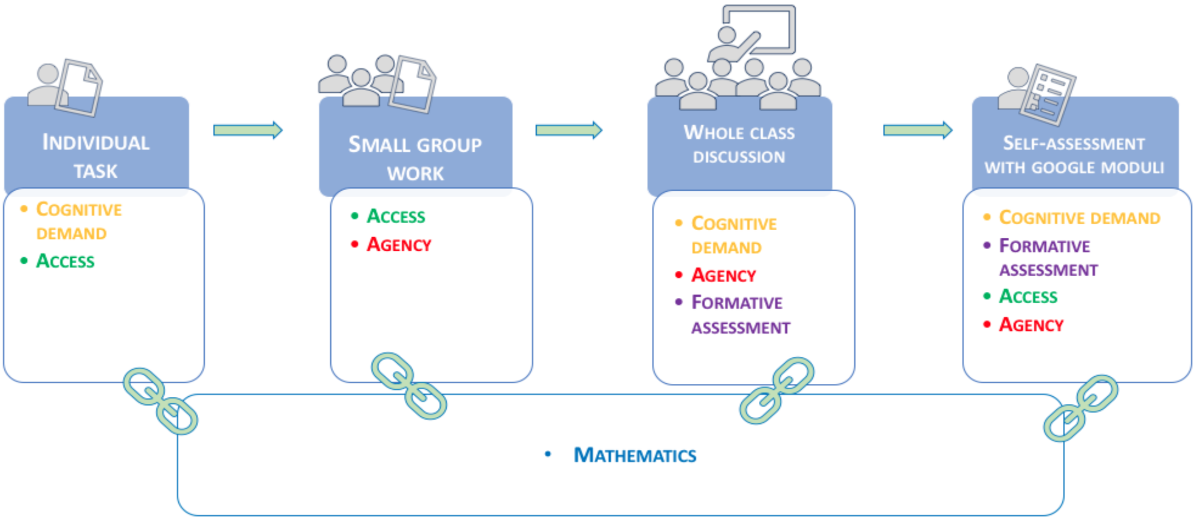


Figure 1: Stages of the activity

Analytical tool: the indicators

To study the actual implementation of formative assessment strategies (Wiliam and Thompson, 2007), we developed specific indicators to detect the *activation* and *realization* of each of them, declined in the specific case of algebraic thinking and argumentation, and indicators for the effective activation of the strategy. The indicators were theoretically set up by tailoring formative assessment strategies to the specific learning goals of the activity, guided by the theoretical frameworks used as references: symbol sense and rational behavior.

Due to space constraints, here we present indicators for FA1 (*clarifying and sharing learning intentions and criteria for success*) and FA2 (*engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding*). These strategies were selected for being the most related to the design of the sequence and to the theoretical frameworks used to frame the learning goals.

The strategies FA1 and FA2 were categorized based on the specific learning goals they addressed: FA1.1 and FA2.1 pertain to algebraic thinking, while FA1.2 and FA2.2 are associated with rational behavior.

Table 1: Indicators for FA1 and FA2

	Indicators for activation	**Indicators for realization**
FA1.1	a. the teacher underlines/makes explicit the importance of using symbols to represent relationships, generalizations and proofs b. the teacher points out that in some cases it is more convenient to abandon symbols in favor of other approaches c. the teacher makes explicit the importance of dealing with the dialectic between manipulation and interpretation of symbols d. the teacher underlines the possibility of creating symbolic expressions, and being able to create them; e. the teacher underlines the importance of selecting but also abandoning or changing a symbolic representation f. the teacher underlines the importance of constantly checking symbol meanings during problem solving g. the teacher underlines that symbols may play different roles	The student shows to be aware of the learning goals and criteria for success concerning algebraic thinking (e.g. mentioning the importance of using algebra to generalize)
FA1.2	a. the teacher underlines the importance of providing explanations b. the teacher clarifies the criteria for a good argumentation c. the teacher promotes a reflection on the epistemic component (correctness) d. the teacher promotes a reflection on the teleologic component (strategy to solve the problem, goal-oriented actions…) e. the teacher promotes a reflection on the communicative component (comprehensibility of the solution, …) f. the teacher promotes a reflection on the role of examples in argumentation	The student shows to be aware of the learning goals and criteria for success concerning argumentation (e.g. recognizing the need to move beyond numeric examples in proving)
FA2.1	The design encompasses activities such as small group work/class discussion/self assessment, aimed at: a. comparing solving strategies and solutions b. reflecting on strengths and weaknesses of the solving strategies /e.g. choice of the formalization) c. making students explicit their solving process	The teacher poses questions aimed at eliciting evidence of student understanding, with reference to algebraic thinking The student provides evidence of his/her understanding, with reference to algebraic thinking.
FA2.2	The design encompasses activities such as small group work/class discussion/self assessment, aimed at: a. comparing argumentations b. reflecting on strengths and weaknesses of the proposed argumentations (e.g. role for the numerical examples in proving) c. making students explicit their choices related to argumentation (e.g. use of specific terms)	The teacher poses questions aimed at eliciting evidence of student understanding and realization of argumentation. The student provides evidence of his/her understanding, with reference to argumentation.

Discussion of results

We employ the aforementioned indicators to detect *activation* and *realization* of formative assessment strategies in the teaching and learning sequence. Due to space limitations, we present only two examples of this analysis.

For strategy FA1 (*clarifying and sharing learning intentions and criteria for success*), we can identify instances of *activation* in the teacher's efforts to clarify the desired learning objectives. The actual *realization* of FA1 can be found in the students' protocols, their contributions during discussions, and answers in the self-assessment questionnaires, reflecting their grasp of the learning objectives. The example we present specifically pertains to FA1.1.

An instance of the *activation* of FA1.1, particularly with reference to indicator *a*, can be seen in the following excerpt (discussion, min. 26.28):

Teacher (T): Well, it's the power of algebra: the ability to generalize and condense infinite numbers into a single symbolic representation. Indeed, that's how it is. 2n: we've written down all those infinite even numbers in a single expression.

Examples of the *realization* of FA1.1, related to indicator *a*, were observed both during the class discussion and in students' responses to self-assessment questionnaires. In the class discussion, students justified their solutions by highlighting the role played by algebra, prompted by the teacher, as shown in the following excerpt (discussion, min. 02.49 - 03.30):

T: How did we go from 6n+9 to 3(2n+3)?
Mary: We performed factorization
T: What was its purpose?
Mary: To find a number, which is 3k. This factorization became k.
T: Matt, would you like to help Mary?
Matt: We modeled what was inside the parentheses by naming it k, and then we got 3k, which means - well, 3k represents all the multiples of 3.

Further, in the self-assessment questionnaires, answering the question “Where does algebra come into play, and how has it helped you?”, students provided responses like the following: "It helped me generalize parts that would have otherwise required infinite examples." (Elia).

Looking at strategy FA2 (*engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding*), we provide an example of its *activation* for FA2.2, involving all three indicators (*a*, *b*, *c*): the strategy focused on comparing protocols with group responses during collective discussions. It allowed the teacher to address a critical aspect during discussions, aiding students in transitioning from arithmetic to algebra: the role of numerical examples in justifying their solutions to the task.

An example of the *teacher's realization* of FA2.2, specifically indicator *b*, can be seen in the teacher's interventions during the whole class discussion (min. 24.12), seeking to understand the challenges arising from students' arguments, especially by providing feedback on where to focus attention, particularly regarding the role of numerical examples:

T: Alright. Why? Florin, if you remember, Florin, why did you provide an example??

Realization of FA2.2 (indicator *b*) concerns both the teacher, highlighting the goal indicated in FA1.2 *f*, and student intervention related to understanding, as referenced in FA1.2 *f*, can be found in the following excerpts of the discussion (min. 07.35 - 08.02):

T: Do you agree that just one example, for instance, a specific example like they took n=0, which is particular as the smallest number, is enough to prove that Luisa is correct, that is, you always get an odd number?

Phil: Since you can insert infinite numbers, well, it's not enough. But one counterexample is enough to refute the theory.

While showcasing these examples, it's important to note that they represent only exemplar instances of the *activation* and *realization* of the formative assessment strategies, concerning exclusively some specific indicators. In our analysis, we observed that these strategies were activated on several occasions and, further than by the teacher, by a multiplicity of students in the class. This provides a measure of the effective implementation of formative assessment concerning FA1 and FA2 in the teaching and learning sequence, in terms of pervasiveness.

Conclusions, limitations and further directions

In the present contribution, we have shown how designing a teaching and learning sequence within the TRU framework, where all the dimensions are shaped around the mathematical learning goals, complemented by specific theoretical references (symbol sense and rational behavior), allows us to formulate a tool for analysis and, consequently, to evaluate the effective implementation of formative assessment strategies (Wiliam & Thompson, 2007) in the classroom. Indeed, the study leads us to conclude that the modes of designing the sequence allowed us, firstly, to identify specific indicators to monitor the activation and realization of formative assessment strategies.

Secondly, from an initial analysis using the formulated indicators, as shown in the examples illustrated, we could assess that the strategies FA1 (*clarifying and sharing learning intentions and criteria for success*) and FA2 (*engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding*) were realized in the classroom. Although, due to space constraints, we were able to show traces of the activation regarding only some indicators and with few examples of realization taken from the discussion and self-assessment questionnaire, the comprehensive analysis highlighted the pervasiveness of formative assessment strategies (strategies were activated on several occasions and different students were involved). In further work, we will illustrate our analysis in relation to the other strategies.

A further direction involves generalizing the analytical tool, thus studying whether the development of contextual indicators for evaluating the effective implementation of formative assessment strategies can be transferred to the analysis of other teaching activities with specific learning goals.

References

Arcavi, A. (1994) Symbol sense: informal sense-making in formal mathematics. *For the Learning of Mathematics*, 14(3), 24–35.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5–31.

Morselli, F., & Boero, P. (2010). Proving as a rational behaviour: Habermas' construct of rationality as a comprehensive frame or the teaching and learning of proof. In V. Durand-Guerrier, S. Soury-Lavergne, & F. Arzarello (Eds.), *Proceedings of CERME 6, 6th Congress of European Research in Mathematics Education*, Lyon (France), 211–220.

Morselli, F., & Quartara, S. (2023). "My mind is getting used to always find a better solution process": Formative assessment and self-regulation in secondary school algebra. In P. Drijvers, C. Csapodi, H. Palmér, K. Gosztonyi, , E. & Kónya (Eds.), *Proceedings of CERME 12, the 12th Congress of the European Society for Research in Mathematics Education*, Alfréd Rényi Institute of Mathematics and ERME, 4020–4028.

Schoenfeld, A. H., & the Teaching for Robust Understanding Project. (2016). *An Introduction to the Teaching for Robust Understanding (TRU) Framework*. Berkeley, CA: Graduate School of Education. Retrieved from http://map.mathshell.org/trumath.php.

Schildkamp, K., van der Kleij, F.M., Heitink, M.C., Kippers, W.B., & Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research*, 103, 101602. https://doi.org/10.1016/j.ijer.2020.101602

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning, 53–82*. Erlbaum. https://doi.org/10.4324/9781315086545

Evaluating the reliability of a framework for mathematical activities using generalizability theory

Ali Bozkurt[1], Mehmet Fatih Özmantar [2] and Sibel Tutan Teskin[3]

[1]Gaziantep University, Faculty of Education, Gaziantep, Turkey; alibozkurt@gantep.edu.tr

[2]Gaziantep University, Faculty of Education, Gaziantep, Turkey; ozmantar@gantep.edu.tr

[3]Ministry of Education, İstanbul, Turkey; sibell27@outlook.com

Abstract

This study investigates the measurement reliability of the Framework for Mathematical Activities (FfMA), developed to assess the quality of activity scripts and their implementations. Utilizing Generalizability Theory, the measurement reliability of scores obtained from the FfMA tool was determined. Data were collected based on a descriptive survey model. In this context, four activity scripts and classroom implementation videos of these texts were requested from each of 20 middle school mathematics teachers. The data were scored independently by three raters using the FfMA tool. The scores obtained from the raters were analyzed using the EDUG 6.1e program. Findings indicate that the measurement reliability of the FfMA tool is considered reliable with a value of 0.78, and it does not fall below 0.70 even in scenarios with the minimum number of raters (2) in D studies. These coefficients suggest that the FfMA tool consistently measures the quality of mathematical activities.

Key words: Mathematical activity script, mathematical activity implementation, generalizability theory, reliability.

Introduction

In mathematics education, activity-based teaching holds the potential to influence students' mathematical success and prepare them to be independent and democratic thinkers (Noreen & Rana, 2019). However, to realize this potential, there is a need for high-quality activity scripts and implementations. A review of the literature reveals various frameworks for assessing the quality of mathematics instruction, such as the Mathematical Quality of Instruction [MQI] (LMTP, 2011), Teaching for Robust Understanding [TRU] (Schoenfeld, 2013), Classroom Assessment Scoring System [CLASS] (Pianta & Hamre, 2009), and Framework for Teaching [F*f*T] (Danielson, 2013). These frameworks are generally designed to determine the overall quality of a lesson and do not provide a detailed evaluation specifically focused on the quality of the instructional activities. On the other hand, studies specifically concerning the mathematical dimensions of activities often limit their focus to particular aspects like cognitive demand, purpose, and materials (Stein & Smith, 1998). However, considering only limited, and often isolated, aspect such as cognitive demand or purpose is insufficient for a comprehensive evaluation, as the quality of a mathematical activity scripts and implementation is influenced by numerous components. In order to fill this gap, Framework for Mathematical Activities (F*f*MA) (Bozkurt et al., 2023) was developed to be used as a feedback tool and to evaluate the quality of activity text and activity applications.

Framework for Mathematical Activities (F*f*MA)

F*f*MA is an assessment tool that can be used to determine the quality of activity scripts and implementations by considering them separately. Based on these assessments, it is intended to be used in such a way as to provide users with feedback on the strengths and improvement areas of the mathematical activity text and the performance of the activity implementation. In order for F*f*MA to have a functional use as an assessment tool, concrete indicators and observable criteria were taken as a basis. In this way, F*f*MA is intended to serve for reliable scoring as well as valid results. The products that F*f*MA evaluates are the activity text and its implementation. The activity script is a concrete tool produced as a document found in various sources or prepared by the teacher himself/herself and has observable qualities. Implementation, on the other hand, occurs in the real classroom environment based on the interaction between the student-teacher-content triad and has observable characteristics. The dimensions and components of F*f*MA are illustrated in Figure 1.

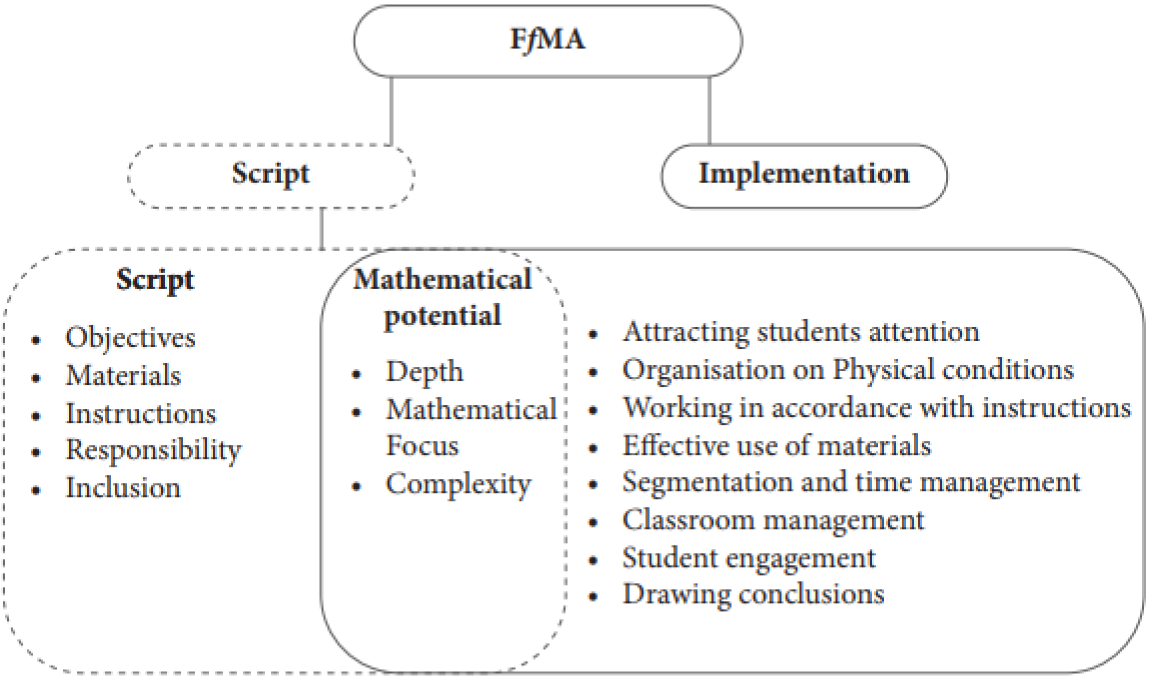


Figure 1. Dimensions and components of F*f*MA (Bozkurt et al, 2023)

As illustrated in Figure 5.1.1, the activity script has a total of 8 components and the implementation has a total of 11 components. Both dimensions, the activity text and the implementation, include an evaluation in terms of mathematical potential. Mathematical potential includes components related to determining the mathematical quality of the activity script and the implementation.

This study examines the measurement reliability of the Framework for Mathematical Activities (F*f*MA) (Bozkurt et al., 2023), which offers a broader perspective for evaluating activity texts and implementations. Utilizing the approach of Generalizability Theory, the study aims to determine the measurement reliability and generalizability of scores derived from the F*f*MA tool.

Method

This study stems from a project which was initiated, funded by TUBITAK (Project Number: #119K773). Spanning two years from 2020 to 2022 (Bozkurt et al., 2022). Within the scope of this study, the reliability study of the F*f*MA developed by the expert researchers in this project was conducted. In the literature, different theories have been developed to test the reliability of measurements obtained from a measurement tool: Generalizability Theory, Classical Test Theory, Multivariate Rash Model, Item Response Theory, etc. These theories differ according to the purpose of use, limitations, and how the measurement results are used (Brennan, 2001). Generalizability studies are organized to determine the source of variability from which measurement errors occur with a single analysis. Generalizability Theory, the chosen method for testing the reliability of F*f*MA, is a statistical theory based on the analysis of variance (ANOVA). It is particularly useful in measurements involving different sources of error, allowing for the estimation of errors stemming from these sources and their interactions. Generalizability studies are divided into two types: G (generalization) studies and D (decision) studies. In G studies, the aim is to determine the variance components that affect reliability, as well as to generalize the measurement taken from the sample to the larger population (Brennan, 2001). D studies provide data for researchers to identify candidates for selection-placement, compare experimental groups, and investigate the relationship between at least two variables (Arterberry et al., 2012).

Data were collected from different schools and different grade levels during one semester in the 2022-2023 academic year. Four activity scripts and video records of their classroom implementations were collected from each of 20 middle school mathematics teachers. Participants were selected through purposive sampling method (Rai & Thapa, 2015). The teachers were asked to design activities or adapt an existing activity and implement it in their classrooms. They were also instructed to video record their implementations in actual settings. The decision regarding which lesson to record and when was left to the discretion of the teachers. The sample size was within the range of similar G-theory models and observational measurements (e.g., minimum 8) (Hill et al., 2012). This approach aligns with the methodological emphasis on the quality and applicability of the data rather than just the quantity, ensuring a more targeted and relevant evaluation of the F*f*MA tool's reliability.

The activity scripts and in-class video recordings of the implementation of these scripts were analyzed by the researchers. At the end of the analyzes, it was seen that in some videos, it was not possible to obtain healthy data on the student-teacher-content triad. These video recordings and the activity scripts used in these implementations were excluded from the evaluation. After this preselection process, 65 activity scripts and corresponding implementation videos were selected for further analysis. The data were independently scored by three raters using the F*f*MA. The components in the F*f*MA are graded on 4 score types (0: Very low; 1: Low; 2: Medium; 3: High). The scores that can be produced in an evaluation with F*f*MA are in the range of 0-24 points as minimum-maximum for the 8-component activity script dimension. For the 11-component activity implementation dimension, the minimum-maximum score range is 0-33 points. The scores obtained from the raters were then analyzed using the EDUG 6.1e program. In this framework, a G (generalizability) study was conducted following the pattern of "*activity script (a) x component interaction effect (c) x rater (r)*" and "*implementation (u) x component interaction effect (c) x rater (r)*". This study involved analyzing variance values for main and interaction effects. Subsequently, a D (decision) study was conducted to calculate the G coefficients for the reliability of the scores.

Findings

Generalizability of F*f*MA's Activity Script Dimension

The G (generalizability) study conducted for evaluating the activity scripts in F*f*MA revealed that the variance component attributed to the activity text source explained 9.7% of the total variance. This indicates that the components in the script tool are capable of distinguishing between different components of the activity script. The variance component estimated for the activity script-component-rater (*a x c x r*) or residual (unobserved or unintended) effect was 0.01736 and this effect explained 47.8% of the total variance. The second-largest source of variance was the interaction between component and rater (*c x r*), accounting for 37.1%. This variance indicates variability in the ratings given by different raters to different components. The interaction between activity script and component (*a x c*), contributing 5.4% to the variance, suggests that the difficulty levels and qualities of the components do not vary significantly from one script to another. Other sources of variance (*c, r, a x r*) were found to contribute zero or near-zero to the total variance, indicating their minimal impact on the overall variability in this context.

The D (decision) studies, which varied the number of raters and components in different scenarios, demonstrated that an increase in the number of raters leads to an improvement in reliability parameters. Based on the relative error variance, the G coefficient was found to be 0.78. In scenarios where the number of raters was three, and the number of components was eight or more, the G coefficients exceeded 0.80. It was also determined that even when the number of raters was reduced to two, the reliability parameters did not fall below 0.70. These findings highlight the robustness of the F*f*MA tool's reliability across different scenarios, emphasizing that even with a reduced number of raters, the tool maintains a satisfactory level of measurement reliability. This consistency in reliability, regardless of the number of raters, underscores the effectiveness of the F*f*MA framework in providing dependable evaluations of mathematical activity scripts.

Generalizability of F*f*MA's Activity Implementation Dimension

The G (generalizability) study focused on the evaluation of activity implementations within the F*f*MA framework revealed that the variance component attributed to the implementation of activities accounted for 5.7% of the total variance. This indicates that the components of the measurement tool are effective in distinguishing between the components of activity implementations. The most significant source of variance was the interaction between implementation, component and rater (*i x c x r*), accounting for 41%. This high level of variance suggests that the scores for the components of the activity implementations varied significantly due to the interaction effect and/or random errors, more than by common effects, from one rater to another and from one component to another. The second highest source of variance was the interaction between component and rater (*c x r*), accounting for 32%. This indicates that there is variability and inconsistency in the ratings given by different raters to different components of activity implementations. Interestingly, the variance component attributable to the raters alone (r) explained 0% of the total variance. This can be interpreted as an indication of the raters providing consistent scores, highlighting their reliability and uniformity in evaluating the activity implementations.

D studies were conducted by creating scenarios with varying numbers of raters and components. According to the results derived from the relative error variance, the G coefficient was found to be 0.78. It was observed that when the number of components increased while the number of raters remained constant, the G coefficient also increased. This suggests that the reliability parameters are affected by the variance in errors, which fluctuate depending on the number of raters and components involved. Even in scenarios where the number of components was reduced, the reliability parameters did not fall below 0.70, even when the number of raters was as low as two. This finding underscores the robustness of the F*f*MA tool's reliability, demonstrating that it maintains a significant level of accuracy and consistency across various testing conditions. It highlights that the F*f*MA is a reliable tool for evaluating mathematical activities, providing dependable results even with variations in the number of raters and components used in the assessment.

Discussion

According to the findings of the study, the generalizability (G) coefficient of the reliability of the measurements obtained with the F*f*MA tool was found to be 0.78. Even in decision studies where the minimum number of raters (two) was used, the reliability coefficient did not fall below 0.70. In the study where the number of raters was 3, G coefficients were above 0.80 in scenarios where the number of components was 8 or more. These coefficients indicate that the F*f*MA tool can produce consistent assessments in different scenarios (Brennan, 2001; Shavelson & Webb, 1991). This demonstrates that F*f*MA measures the quality of mathematical activities in a consistent and reliable way.

Another finding of the study is the residual variance effect, which has the highest source of variance. The raters' contribution to the shared variance is zero, indicating their consistency in scoring. This finding shows that the quality of the activities varies from component to component and from rater to rater, and is more affected by the common effect and/or random errors. A similar finding was found in the study conducted by Solano-Flores (2006). In this study, G theory was used to estimate the amount of measurement error in the sources of variance of the instrument dealing with psychometric approaches to testing English language learners. Two groups were measured by giving questions in different dialects. The largest measurement error observed was due to the interaction of student, item and code (residual variance/random variance). The high residual variance may be due to various sources of variance not included in the design as well as the interaction effect (Uzun et al., 2018). Moreover, the decision studies reveal that reducing the number of raters does not significantly compromise the reliability of the tool. In fact, it was found that as the number of raters and components increases, so does the reliability coefficients. This finding is significant as it demonstrates the robustness of the F*f*MA tool in different evaluation contexts, ensuring that it remains a reliable measure for assessing the quality of mathematical activities, even with variations in the number of raters and components used.

Some of the coefficients obtained from the research were found to be low, which could be attributed to the closeness of the scores derived from the activities included in the study. This is because G coefficients are negatively affected by group homogeneity. According to Generalizability Theory, G and phi coefficients are calculated by relating universe variance to observed variance (Brennan, 2001). In other words, as the similarity between groups increases, the variance value decreases, which in turn could explain why the calculated G coefficients are not found to be high. A similar situation was observed in the study by Ozbası and Arcagok (2021), where they examined students' projects in a fully crossed design (*j x p x i*) within the framework of generalizability theory, considering jurors, projects, and items. In their research, the homogeneity among the projects led to a decrease in variance, thus affecting the generalizability coefficients in a similar manner. This underlines the impact of group homogeneity on the reliability measures in generalizability studies, suggesting that while the FfMA tool is reliable, the nature of the activities and the scoring patterns can influence the overall variance and, consequently, the generalizability coefficients.

As a result, considering the studies conducted, the reliability of this measure obtained from F*f*MA is acceptable. As a result of the decision studies, it was observed that the G coefficient increased more in scenarios where the number of raters was increased while the number of components was kept constant. This study contributes to the literature in terms of determining the factors affecting the reliability of the scores obtained from the evaluation of the quality of activity design and implementation in mathematics education through G theory. Based on our findings, it is suggested to provide teachers with guidance on how to evaluate and score mathematical activities. This can include training or resources that clarify the assessment process within the framework of the F*f*MA tool. Furthermore, to enhance the reliability of assessments, it could be beneficial to involve peer and expert evaluations of activity scripts and implementations across different student groups. By allowing for a diverse range of assessments, it would be possible to analyze the measurement results within the framework of Generalizability Theory more comprehensively. Such initiatives could not only improve the accuracy of evaluations but also provide valuable insights into the effectiveness of mathematical activities in diverse educational settings.

Acknowledgments

This research project was supported by Scientific and Technological Research Council of Turkey (TÜBİTAK 1001 no: 119K773)

References

Arterberry, B. J., Martens, M. P., Cadigan, J. M. & Smith, A. E. (2012). Assessing the dependability of drinking motives via generalizability theory. *Measurement and Evaluation in Counseling and Development, 45(4), 292-302.*

Baykul, Y. (2000). Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulamasi. Ankara: ÖSYM yayınları.

Bozkurt, A., Özmantar M.F., Agaç, G., & Güzel, M. (2022). Developing an Evaluation and Feedback Tool for Determining the Quality of Mathematical Task and Implementation. *Project Report, TÜBİTAK 1001 119K773.*

Bozkurt, A., Özmantar M. F., Agaç, G. & Güzel, M. (2023). *A Framework for Evaluating Design and Implementation of Activities for Mathematics Instruction.* Ankara: Pegem Academy.

Brennan, R. L. (2001). *Generalizability theory*. USA: Springer-Verlag New York Inc.

Danielson, C. (2013). *The Framework for teaching evaluation instrument,* 2013 edition. The Danielson Group.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56-64.

Karasar, N. (2008). *Scientific Research Method.* Ankara: Nobel Publishing House.

LMTP (Learning Mathematics for Teaching Project), (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, 14, 25-47

Noreen, R., & Rana, A.M.K. (2019). Activity-Based teaching versus traditional method of teaching in mathematics at elementary level. *Bulletin of Education and Research*, *41*(2), 145-159.

Ozbası, D. & Arcagok, S. (2021). Examining student projects with Generalizability Theory, *Journal of Theory and Practice in Education, 17*(2), 69-78. doi: 10.17244/eku.1024532

Pianta, R.C. & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119.

Rai, N., & Thapa, B. (2015). A study on purposive sampling method in research. *Kathmandu: Kathmandu School of Law*, *5*(1), 8-15.

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English language learners. *Teachers College Record*, *108*(11), 2354-2379.

Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM*,*45*(4), 607–621.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory a primer*. California: Sage Publications.

Stein, M. K., & Smith, M. S. (1998). Mathematical tasks as a framework for reflection: From research to practice. *Mathematics teaching in the middle school*, *3*(4), 268-275.

Uzun, N. B., Alıcı, D., & Aktas, M. (2018). Reliability of the analytic rubric and checklist for the assessment of story writing skills: G and decision study in generalizability theory. *European Journal of Educational Research, 8(1), 169-180.*

Expert design and implementation of effective classroom discussions for formative assessment

Annalisa Cusi[1], Francesca Morselli[2] and Cristina Sabena[3]

[1]Sapienza Università di Roma, Italy; annalisa.cusi@uniroma1.it

[2]University of Genova, Italy; morselli@dima.unige.it

[2]University of Torino, Italy; cristina.sabena@unito.it

We present the first results of a long-term study aimed at characterizing an expert design and implementation of effective classroom discussions for formative assessment. For the analysis of the data collected in this study we combine the use of three different theoretical constructs concerning: the expert teacher's roles during classroom discussions; shared attention; formative assessment key-strategies. The presented results concern, on one side, the expert's use of a specific digital technology (an interactive whiteboard) to empower specific teacher's roles to promote shared attention and, on the other side, the effects of the empowered teacher's roles in the activation of specific formative assessment strategies.

Keywords: formative assessment, classroom discussion, teacher's roles, shared attention.

Introduction and theoretical background

In this contribution we present the first results of a long-term study aimed at characterizing an *expert* design and implementation of *effective* classroom discussions for formative assessment (FA).

We conceptualize teacher's expertise by referring to Mason and Spence (1999). Specifically, in our perspective, an *expert* design and implementation is realized by a teacher who knows-to as well as knows-how "to create suitable conditions and then to direct student attention effectively" (p. 147). This is linked to the teacher's awareness of the fact that "it is so vital for students to have the opportunity to be in the presence of someone who is aware of the awarenesses that constitute their mathematical 'seeing'" (p. 151).

In our perspective, classroom discussions are *effective* for FA if they support the activation of *FA key-strategies* (Wiliam & Thompson, 2007): (A) clarifying and sharing learning intentions and success criteria; (B) engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding; (C) providing feedback that moves learners forward; (D) activating students as instructional resources for one another; and (E) activating students as the owners of their own learning.

In line with Mason and Spence's (1999) ideas, our hypothesis is that promoting *shared attention* may foster fruitful FA processes. This is in tune with recent studies, developed in the field of mathematics education, on the role of moments of joint attention in fostering students' acquisition of a culturally appropriate meaning of mathematical objects (Shvarts, 2018; Salminen-Saari et al., 2021). To define shared attention, scholars (Shteynberg, 2015, Siposova and Carpenter, 2019, Fredriksson, 2022) stress on the crucial difference between the social-cognitive processes that take place when people act as detached observers of each other (third-person perspective), and the processes in which individuals interact by adopting an engaged attitude towards each other (second-person perspective).

Fredriksson (2022) emphasises that, when shared attention is realised, "a first-person perspective may develop into a we-perspective in which it is not an I, but "a we", that is attuned to the world" (p. 115). In shared attention, two different beings "find the same attunement with the world" (p. 114) and acknowledge the commonality of their world.

To characterize the expert design and implementation of effective classroom discussions, we have analysed a large amount of data collected during the FaSMEd Project (Cusi, Morselli & Sabena, 2017), during which we carried out teaching experiments focused on the use of connected classroom technologies and Interactive Whiteboards (IWB) to support teachers' FA practices.

The data analysis has been developed by referring to a theoretical construct useful for interpreting and analysing teachers' roles, namely the *Model of Aware and Effective Attitudes and Behaviours, MAEAB* (Cusi & Malara, 2013, 2016). The MAEAB construct identifies two main groups of roles that an expert teacher intentionally plays, during a classroom discussion, with the main aim of "making thinking visible" and of stimulating metacognitive reflections. The first group of roles are those that the teacher plays in order to pose him/herself as a model by making visible the hidden thinking, the aims, the meaning of the strategies, and the interpretation of results when facing problems: (1) investigating subject and constituent part of the class; (2) practical-strategic guide; (3) activator of interpretative processes; (4) activator of anticipating thoughts. The second group includes the roles that the teacher plays when he/she stimulates metacognitive reflections to help students become aware of the meaning of the realized activities and of the learning processes themselves: (5) guide in fostering a harmonized balance between the syntactical and the semantic level; (6) reflective guide in the identification of effective practical-strategic models; (c) activator of reflective attitudes and meta-cognitive acts.

Research method

Within the FaSMEd project, the teaching experiments took place in 36 classes encompassing students from 4th to 7th grade, across two consecutive school years (2014–15 and 2015–16), in three school clusters in north-western Italy. We collaborated with 20 teachers to collect approximately 450 hours of classroom sessions. During the teaching experiments, the role of the expert was played by a researcher, one of the authors, in line with the Italian paradigm of research for innovation (Arzarello & Bartolini Bussi, 1998), which theorizes the elimination of the classical distinction between observer and observed (on one side, the class, including the teacher, and, on the other side, the researcher).

We collected lesson's video recordings, observers' field notes and students' written answers. Video recordings and their transcripts form the data corpus for the part of the study documented in this paper. The transcripts were analysed separately by the three authors. Non-converging elements of the analysis were discussed further so as to reach an agreement. We combined the use of the aforementioned theoretical constructs to study how the expert teacher designs and implements classroom discussions through the support of an IWB to empower the MAEAB's roles by promoting shared attention that fosters fruitful FA key-strategies. More specifically, the expert's interventions:

- were analysed according to the MAEAB construct (Cusi & Malara, 2013, 2016);
- were related to the foci of shared attention (Fredriksson, 2022) that they aimed to promote;

- were linked to their effects in terms of the FA key-strategies (Wiliam & Thompson, 2007) activated by means of these interventions.

In this 4-pages presentation we confine ourselves to outline the main results for the analysis. The analysis of a paradigmatic example will be added in the oral communication.

Results and discussion

The first set of results concerns the expert's use of the IWB to empower specific MAEAB roles to promote shared attention. We found that specific uses of the IWB and other specific expert's actions empowered most of the MAEAB roles: zooming-out and/or scrolling from top to bottom to provide an overall view of the groups' answers; zooming-in to focus on particular answers; scrolling up to focus on elements of the given task; inviting one student to come to the IWB to comment on his answer focusing on both the answer and the text of the task; standing in front of the IWB, reformulating a student's discourse and repeating her/his gestures. These uses and actions promote shared attention on different foci: the task and its elements; the overall distribution of students' answers (collective product); a specific written answer (single product) and its characteristics; the approach taken by a student to solve the task (past thinking process); in current approach and reflection on the task (present thinking process).

This analysis highlighted elements of synergy between the shared attention construct and the activation of the roles introduced by the MAEAB construct:

- the shift from the "I-perspective" to the "we-perspective" (which is an indicator of the role of *investigating subject and constituent part of the class*),
- the intentional communication about a common object of attention (i.e. a representation in the case of the role of activator of interpreting processes, a strategy or an argument in the case of the role of *activator of reflective attitudes and metacognitive acts*, the thinking processes of a student or of the teacher in the case of the roles of *reflective guide* and *practical-strategic guide*),
- the focus on metacognitive processes.

The second set of results concerns the effects of the empowered MAEAB roles in the activation of specific FA strategies. The roles of activator of reflective attitudes and metacognitive acts and of reflective guide, contribute to the promotion of specific FA strategies. For instance, teachers can encourage shared attention on a subset of responses, fostering peer assessment among students and thereby promoting FA strategy D. Additionally, students are encouraged to offer feedback to one another, thus realizing FA strategy C. Teachers may also prompt meta-level reflections on provided answers or the reasoning behind them, encouraging students' self-assessment and thus promoting FA strategy E.

In the same way, we found examples of links between the role of guide in fostering a harmonized balance between the syntactical and the semantic level and FA strategy A, and the roles of operative-strategic guide and activator of interpretative processes and FA strategy E.

This study has two implications: (1) at the theoretical level, the study shows the effectiveness of combining the MAEAB construct and the construct of shared attention to gain insights into the ways in which the expert teacher may promote FA during classroom discussions; (2) at the pragmatic level, this combination could provide a tool for teacher professional development aimed at promoting teachers' autonomous design and implementation of effective classroom discussions for FA.

References

Arzarello, F., & Bartolini Bussi, M. (1998). Italian trends of research in mathematics education: A national case study in the international perspective. In J. Kilpatrick & A. Sierpinska (Eds.), *Mathematics education as a research domain: The search for identity* (vol. 2, pp. 243–262). Kluwer.

Cusi, A., & Malara, N.A. (2013). A theoretical construct to analyze the teacher's role during introductory activities to algebraic modelling. In B. Ubuz, C. Haser & M.A. Mariotti (Eds.), *Proceedings of Cerme 8* (pp. 3015–3024). Middle East Technical University and ERME.

Cusi, A., & Malara, N.A. (2016). The Intertwining of Theory and Practice: Influences on Ways of Teaching and Teachers' Education. In L. English, & D. Kirshner (Eds.), *Handbook of International Research in Mathematics Education 3rd Edition* (pp. 504–522). Taylor & Francis.

Cusi, A., Morselli, F., & Sabena, C. (2017A). Promoting Formative Assessment in a Connected Classroom Environment: Design and Implementation of Digital Resources. *ZDM - Mathematics Education*, *49*(5), 755–767.

Fredriksson, A. (2022). *A Phenomenology of Attention and the Unfamiliar Encounters with the Unknown*. Palgrave Macmillan.

Mason, J. & Spence, M. (1999). Beyond mere knowledge of mathematics: the importance of knowing-to act in the moment. *Educational Studies in Mathematics*, *38*, 135–161.

Salminen-Saari, J. F. A., Garcia Moreno-Esteva, E., Haataja, E., Toivanen, M., Hannula, M. S., & Laine, A. (2021). Phases of collaborative mathematical problem solving and joint attention: a case study utilizing mobile gaze tracking. *ZDM - Mathematics Education*, *53*, 771–784.

Shteynberg, G. (2015). Shared attention. *Perspectives on Psychological Science*, *10*(5) 579–590.

Shvarts, A. (2018). Joint attention in resolving the ambiguity of different presentations: a dual eye-tracking study of the teaching-learning process. In N. Presmeg, L. Radford, W.-M. Roth & G. Kadunz (Eds.), *Signs of Signification. Semiotics in mathematics education research* (pp. 73-102). ICME-13 Monographs. Springer.

Siposova, B., & Carpenter, M. (2019). A New Look at Joint Attention and Common Knowledge. *Cognition*, *189*, 260–274.

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Lawrence Erlbaum Associates.

Enacting assessment accommodations in an inclusive formative classroom practice: The case of color-coding

Angeliki Dikarou[1] and Chrissavgi Triantafillou[2]

[1]National and Kapodistrian University of Athens; angeliki_ad@hotmail.com

[2]National and Kapodistrian University of Athens; chrtriantaf@math.uoa.gr

This study explores how a Special Education Teacher in Mathematics (SETM) implements the assessment accommodation of color-coding in a Parallel Support setting. SETM's goal is to support a student with learning disabilities in a grade 8 mathematics classroom. We view formative assessment as a unified classroom practice that involves teachers' actions and students' responses to these actions. The results indicate that the main SETM's actions while enacting color-coding accommodation are a*) repeating and extending student's short answers; b) asking student to justify his responses; c) evaluating and validating student's correct responses; d) honoring student's contribution by maintaining his mathematical idea and e) creating a positive and engaging learning environment by frequently rewarding student's responses.*

Keywords: Inclusive formative classroom practices, assessment accommodations, color-coding, special education teacher's actions.

Introduction

The potential of using formative assessment in mathematics classrooms to raise students' learning is well documented in many studies (Andersson, 2020; Heritage & Wylie, 2018). Even though, many studies argue that adapting formative assessment practices could be more effective and inclusive for specific group of learners such as learners with autism (e.g., Ravet, 2013), there is a limited number of studies that explore formative assessment in mathematics classrooms from a special education perspective.

In special education settings, teachers usually implement assessment accommodations to support students with special learning needs (Maccini & Gagnon, 2000). Assessment accommodations are changes made to an assessment procedure (e.g., scheduling, timing, task presentation), that aim to remove barriers and allow students to fully demonstrate their competencies and their abilities (Elliott et al., 1998). Maccini and Gagnon (2000) determined the type of assessment accommodations that special and general education teachers reported while enacting assessment practices. These types of accommodations may include visual tools for task presentation; reference materials such as cue cards or charts of strategy steps; or time extensions on tests.

The current study explores how a Special Education Teacher in Mathematics (SETM) implements a specific type of visual tools such as color-coding in order to support a student with autism. The research question (RQ) is: What are SETM's teaching actions while enacting the assessment accommodation of color-coding in a formative assessment practice?

Color-coding is the use of colors to represent data values on a task. This means that every data value is associated with exactly one color, and vice versa i.e. every color represents a fixed range of data values (Tominski et al., 2008).

Theoretical Background

Socio-cultural perspective

We adopt a socio-cultural perspective since we view teaching and learning as a joint labor process where teachers and students are laboring together to produce knowledge (Radford, 2014). Thus, it is important to explore teachers' in-the-moment responses to student mathematical contribution. In this study, we analyze SETM's actions through the Teacher Response Coding (TRC) framework (Van Zoest et al., 2022). Some of the actions included in this framework are *allow* (creates an open space for interaction); *check-in* (elicits student's self-assessment or understanding); *clarify* (asks the student to make more precise answer); *justify* (gives the student the opportunity to reason on his mathematical idea) and *evaluate* or *validate* students' responses. Another important aspect of the TRC framework is the degree to which the teachers' response aligns with students' ideas and mathematical contributions.

Inclusive formative assessment practice

Practice in a classroom, is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about next steps in instruction (Black & Wiliam, 2009). Thus, all studies share the defining characteristic of formative assessment: agents in the classroom collect evidence of students' learning, and based on this information, adjust their teaching and/or learning (Andersson & Palm, 2017). In this study we use the term of 'inclusive formative assessment' (Andersson, 2020, p. 75) where inclusion means that students' diversity and differences are seen as something natural and valuable. Based on this perspective, students with mild learning disorders are now taught in mainstream classrooms and not in special units and schools. To achieve inclusive formative assessment is a challenging issue. Ravet (2013, p. 961) argues that "inclusive formative assessment can be more successful when teachers abstract themselves from the straitjacket of normative thinking about learning, in order to understand the minds of students who function differently."

In this study, we view formative assessment as a unified practice that involves teachers' actions and students' responses to these actions. We analyze SETM's actions while enacting a specific assessment accommodation in a Parallel Support environment as well as the outcome of these actions on students' learning.

Literature review

A limited number of studies suggest empirically validated approaches for assessing students with learning disabilities in mathematics classrooms. Tay and Kee (2019) study mainstream teachers' effective questioning and feedback in primary and secondary math and science classrooms that include high-functioning students with autism spectrum disorder (ASD). They identified three important characteristics of effective questioning and feedback for these students: *addressing students' cognitive needs of* (e.g., precise and direct questions); *taking into consideration their socio-emotional needs* (e.g., affirmative feedback); and *using of supporting structures* (e.g., visual cues). Andersson (2020) documented 39 special education teachers' views while implementing formative assessment practices in mathematics classrooms. Participants, referred to the potential of formative assessment for students with learning disabilities as well as to the challenges they faced, while trying to adjust the learning environment according to students' individual needs.

Methodology

The Greek educational system

The Greek educational system, based on the current legislation (Law 3699 of 2008, article 6) provides inclusive teaching support programs, such as the Parallel Support (PS) program, for students with learning disabilities (e.g., students with Autism Spectrum Disorder (ASD) or students with intellectual disability). Parallel Support is a co-teaching program where two teachers, a general education teacher and a special education teacher share the instruction for a single student in a single classroom setting. Mavropalias and Anastasiou (2016) explored the features of the Greek co-teaching model of Parallel Support (PS) in several Greek educational regions. Their study revealed that the PS program is similar to the One Teach, One Assist approach where the special education teacher typically sits next to the student with a disability, while the general education teacher delivers the lesson in the traditionally arranged classroom setting. The Special education teachers provide individualized support for these students during lessons in a regular classroom, not only to support them to follow the general education system curriculum, but also to reach their educational needs.

The context of the study

The research was carried out in a general education junior high school, during the 2023-24 school year, where one of the researchers works as SETM in a PS program and is responsible for implementing individualized instruction in mathematics, for students with learning disabilities. In this study the research data concerns one of these students, who attends the 8th grade mathematic classes, with an ASD diagnosis. For the needs of this research ethical issues were taking into consideration. SETM from the beginning of the school year had knowledge about important characteristics of the student's learning profile, through the official written diagnosis. This diagnosis, among others, provided useful instructive suggestions that were estimated to favor student's understanding. Specifically, some of these suggestions were that SETM should conduct a combined review of acquired knowledge and implement applications of mathematical skills; systematically pursue student's understanding of mathematical concepts; use information coding (e.g., acronyms, highlighting or color-coding) and positive reinforcement by rewarding student's effort.

Research data and data analysis

Research data is drawn from the research diary kept by the SETM, concerning her everyday actions as special education teacher. Data derived from the research diary included written notes of her daily schedule; photographic material from the student's notebook; short indicative dialogues with the students and the classroom teacher, written on field notes during the lesson or during the breaks; her teaching goals and her reflections after the lesson enactment; short reports/updates about the students' learning progress; information about students' daily homework tasks and the difficulties they faced. It also included short discussions with general education teachers, concerning assessment tools. The above information supports SETM to gain a better perception of the student's learning profile, learning needs as well as types of assessment accommodations that appeared to have a positive learning outcome for them.

The analysis of research data was carried out in three steps. Step 1: We traced all episodes in SETM's Research Diary where she implemented the specific assessment accommodation of color-coding in a number of lessons. Step 2: Two episodes were selected in which, the color-coding played a significant role in a problem-solving process. Step 3: In these episodes, SETM's actions were analyzed through the TRC framework (Van Zoest et al., 2022).

Results

The following two episodes are from the same teaching hour and the teaching chapter covers the calculation of the area of known geometric shapes. The color-coding method was used during the problem-solving process.

1st episode

The 1st episode was concerning a textbook homework task (Task 1) assigned by the general education teacher to all students. The problem was asking students to calculate the area of the two roads (brown rectangles) and the lawn (green areas) as presented in Figure 1.

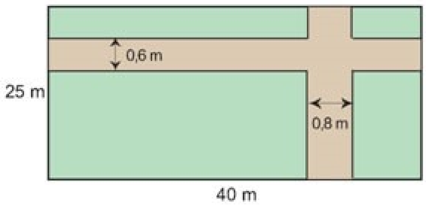


Figure 1: The textbook task (Vlamos et al., 2021, p. 125)

At first SETM checked student's notebook to make sure that homework was done. From this check, SETM realized that there was a mistake in student's following response: 1000 - (24 + 20) = 1000 – 44 + 0.48 = 956.48. In the area calculation solution, student added 0.48 (the crossroad area) without any explanation. This prompted SETM to ask student to justify his answer. The student seemed to face difficulties in justifying his answers. SETM decided to copy the shape on student's notebook by using different colors for each road and the lawn area (Figure 2). Specifically, she took the following color-coding steps as appears in her new version Figure 1 (see Figure 2). Step 1: SETM drew the main rectangle (25W x 40L) and defined with green color all the areas that supposed to be planted with lawn, keeping the book's initial color coding, specifying that this was the area to be calculated. Step 2: Orange color was used to mark the horizontal road. Even though student pointed out correctly the largest and the smallest dimension of the orange rectangle, seemed to struggle when was asked to calculate this area. Step 3: Purple color was used for the vertical road. The main purpose of color-coding was to make the magenta rectangle visible to the student. Then, SETM asked again the student to calculate this area where similar difficulties appeared once more. Then SETM started to discuss with the student about the colors that appeared in the final design.

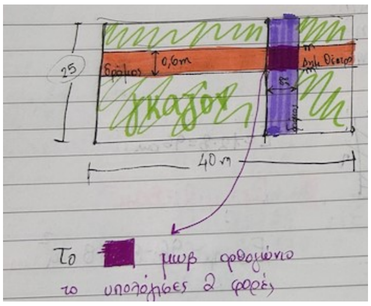


Figure 2: Modifying the textbook task by using color-coding accommodation

1 SETM: So far, we have calculated the area of the orange road and the area of the purple road. Can you see what is happening with that little rectangle with the different color? Why do you think the color changes?
2 Student: Because is the purple above the orange. It has two colors [purple and orange]. This [rectangle] is part of the two roads!
3 SETM: Very nice! So practically what does this mean for us? When we calculated the orange area, we calculated the area of the little magenta rectangle for the 1st time, but we calculated the area of this exact same rectangle for the 2nd time when we calculated the area of the purple road. So, in the end it's like we have calculated magenta's rectangle area twice.

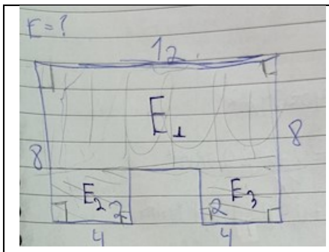
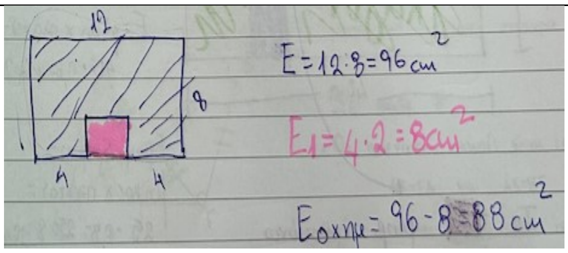
SETM *summarizes* what was done so far and asks student to focus on the color change that appears in the Figure 2 (Line 1). SETM *allows* student to respond and asks him to *justify* through the two colors. Student realized that the magenta area appeared twice in calculations during the solution (Line 2). Student came to this conclusion through the observation that the different color in the crossroad rectangle is due to the overlapping of the two colors. Then, SETM gives supportive feedback to student for the observation made, *repeats* and *extends* the whole solution process in detail (Line 3) and finally she relates the solution with the color-coding accommodation and presents it in mathematical terms by *evaluating* and *validating* student's response. Furthermore, she *honors* and *rewards* student's mathematical contribution.

2nd Episode

The 2nd episode concerned a geometrical problem that the general education teacher gave to the students as homework. This geometric task (Task 2) referred to the calculation of the area of a specific geometric shape. Student drew the shape in the notebook and solved the problem as appears in Figure 3. Then, the general education teacher asked students to provide an alternative solution. Student could not think of a different way to solve the task. The SETM decided to use color-coding to support student to identify another way to solve the problem.

SETM redesigned the geometric shape and filled with pink color the inner rectangle appeared, as shown in Figure 4.

1 SETM: So, can you tell me what is the geometric shape that you see here?
2 Student: A large rectangle.
3 SETM: What's the area of this large rectangle?
4 Student: 12x8

	 12 8 4 4 E=12·8=96cm² E1=4·2=8cm² Eσχημ=96-8=88cm²
Figure 3: Students' initial solution of Task 2	Figure 4: The modification of Task 2, indicating the alternative solution and the use of color-coding accommodation

In Lines 1, SETM starts with *check-in* student's understanding of the geometrical figure. Student possibly identifies two rectangles in the drawing, the large one and the small one colored in pink (Line 2). Then, SETM asks student to *clarify* his answer and make the relevant calculations. SETM wrote "E1 = …" on the notebook to *allow* him to move to the next step i.e. to calculate the area of the small pink rectangle.

5 SETM: So, can you tell me what is the length and the width of this little pink rectangle that was formed?
6 Student: 4 [points out the length of the rectangle].
7 SETM: Oh, nice! And how did you find it?
8 Student: At the left and at the right is also 4. There are 3 pieces that makes us 12.
9 SETM: That is because our shape is rectangle, so the opposite sides are… [let the student finish her argument]
10 Student: Equal.
11 SETM: Perfect! So, you made the calculations 4+4=8 and then 12-8=4. And what a nice observation that in this case we have indeed 3 equal parts of 4, that makes us 12. What about the other dimension? Look at the shape above and then calculate the area of this pink rectangle.
12 Student: 2. So the area is 2x4.
13 SETM: Finally, to calculate this area [outlines the shape with purple lines], what shall we do?
14 Student: 96-8.
15 SETM: Very nice! So, from the area of the large rectangle we will subtract the area of the small one.

In line 5 SETM starts focusing on the small pink rectangle, asking him to *name* the dimensions of the small rectangle. In lines 6 to 8 student responds correctly and SETM asks him to *justify* his responses. It seems that color-coding facilitated student to reach the conclusion that the length is divided into 3 equal pieces (Line 8). SETM *validates* this response while mentioning the relevant theory (Line 9). In line 11 SETM provides positive feedback to the student and then *extends* and *repeats* the mathematical process leading to student's correct answers in lines 6, 8 and 10. Finally, in lines 12 and 14 student gives the correct numerical solution to the problem while SETM *validates* student's answer and keeps *honoring* and *rewarding* student's mathematical contribution (Line 15).

Conclusions

In this paper we explore SETM's actions while enacting a color-coding assessment accommodation in a PS program in a Grade 8 mathematics classroom. These actions constitute parts of an inclusive formative assessment classroom practice. The outcome of SETM's actions was leading to student's understanding. The main SETM's actions while enacting color-coding accommodation are a) repeating and extending student's short answers; b) asking student to justify his responses; c) evaluating and validating student's correct responses by underlining the mathematical reasoning behind these answers; d) honoring student's contribution by maintaining his mathematical idea. In this way the student could easily follow the whole discussion (Van Zoest et al., 2022); and e) creating a positive and engaging learning environment (Hill & Seah, 2023) by frequently rewarding student's responses and focused observations on the color-coding task presentation. SETM's actions seems to satisfy aspects of Tay and Key (2019) effective questioning and feedback. Specifically, SETM poses direct questions when she was asking student to *name* the dimensions of the rectangle; she provides affirmative and constructive feedback; and uses visual cues in color-coded form.

From our perspective, it is the dynamic nature of formative assessment, which makes this process challenging for special education teachers, as it requires continuous adjustments to create the appropriate inclusive conditions for students with learning disabilities (Andersson, 2020). The assessment accommodations -such as color-coding- create opportunities for the learners to demonstrate their mathematical competence and open ways to assess their abilities and not their disabilities (Elliott et al., 1998). Furthermore, despite the institutional limitations that the Greek educational system poses to SETM's teaching activities, as addressed by Mavropalias and Anastasiou (2016), SETM managed to overcome these limitations and to deliver a positive outcome for the student she is responsible for.

Finally, the limitations of this study include, the limited number of participants, one SETM and one student, and the small and by convenience selected sample of episodes does not allow us to generalize our conclusions. More research is needed concerning the enactment of assessment accommodation in inclusive formative assessment practices.

References

Andersson, C. (2020). Formative assessment–from the view of special education teachers in mathematics. *Nordic Studies on Mathematics Education*, *25*(3-4), 73-93.

Andersson, C., & Palm, T. (2017). Characteristics of improved formative assessment practice. *Education Inquiry*, *8*(2), 104-122. https://doi.org/10.1080/20004508.2016.1275185

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability 21*(1), 5-31. https://doi.org/10.1007/s11092-008-9068-5

Elliott, J., Ysseldyke, J., Thurlow, M., & Erickson, R. (1998). What about assessment and accountability? Practical implications for educators. *Teaching Exceptional Children*, *31*(1), 20-27. https://doi.org/10.1177/004005999803100103

Heritage, M., & Wylie, C. (2018). Reaping the benefits of assessment for learning: Achievement, identity, and equity. *ZDM - The International Journal on Mathematics Education, 50*(4), 729-741. https://doi.org/10.1007/s11858-018-0943-3

Hill, J. L., & Seah, W. T. (2023). Student values and wellbeing in mathematics education: perspectives of Chinese primary students. *ZDM–Mathematics Education*, 55(2), 385-398. https://doi.org/10.1007/s11858-022-01418-7

Law 3699. (2008). *Special education and education of people with disability or special educational needs*. Athens: FEK 199A/October 2nd, 2008.

Maccini, P., & Gagnon, J. C. (2000). Best practices for teaching mathematics to secondary students with special needs. *Focus on exceptional children*, 32(5).

Mavropalias, T., & Anastasiou, D. (2016). What does the Greek model of parallel support have to say about co-teaching?. *Teaching and Teacher Education*, *60*, 224-233. https://doi.org/10.1016/j.tate.2016.08.014

Radford, L. (2014). On teachers and students: An ethical cultural-historical perspective. In Nicol, C., Oesterle, S., Liljedahl, P., & Allan, D. (Eds.) *Proceedings of the Joint Meeting of PME 38 and PME-NA 36,* Vol. 38, pp. 1-20. Vancouver, Canada.

Ravet, J. (2013). Delving deeper into the black box: formative assessment, inclusion and learners on the autism spectrum. *International Journal of Inclusive Education*, *17*(9), 948-964. https://doi.org/10.1080/13603116.2012.719552

Tay, H. Y., & Kee, K. N. N. (2019). Effective questioning and feedback for learners with autism in an inclusive classroom. *Cogent Education*, 6(1), 1634920. https://doi.org/10.1080/2331186X.2019.1634920

Tominski, C., Fuchs, G., & Schumann, H. (2008). Task-driven color coding. In *12th International Conference Information Visualisation* (pp. 373-380). IEEE.

Van Zoest, L. R., Peterson, B. E., Rougée, A. O., Stockero, S. L., Leatham, K. R., & Freeburn, B. (2022). Conceptualizing important facets of teacher responses to student mathematical thinking. *International Journal of Mathematical Education in Science and Technology, 53*(10), 2583-2608. https://doi.org/10.1080/0020739X.2021.1895341

Vlamos, P., Droutsos, P., Presvis, G., & Rekoumis, K. (2021). *Grade 8 Mathematics textbook*. OECD. Athens, Greece.

Large language models as formative assessment and feedback tools? – A systematic report

Frederik Dilling

University of Siegen, Germany; dilling@mathematik.uni-siegen.de

This paper discusses the use of large language models (LLMs) for formative assessment and feedback in mathematics education. First, a brief introduction to the research on LLMs in mathematics education is given. Subsequently, the LLM ChatGPT 4.0 is systematically evaluated with regard to the aspects 1) Input and localization, 2) Assessment-quality, 3) Content and form of feedback, and 4) Adaptivity and receiver of feedback. It is shown that ChatGPT has the potential to provide meaningful feedback on mathematical work, but that its use is associated with a number of challenges.

Keywords: Artificial intelligence, ChatGPT, Digital assessment and feedback, Formative assessment, Large language models

Introduction

Artificial intelligence (AI) is currently a highly debated topic. At the latest with the free publication of GPT-3 in November 2022, the discussion has arrived in society. The opportunities and challenges for the educational sector were also quickly addressed. For example, a study by Kung et al. (2023), which found that ChatGPT could pass the three-part American medical licence test (USMLE) without further training, received a great deal of media attention. Opportunities and challenges were also investigated in the field of mathematics education. Wardat et al. (2023) conducted interviews with students and teachers and found that ChatGPT is generally perceived as a useful educational tool, but that it does pose some challenges (e.g. development of misconceptions). Other authors, however, analyzed ChatGPT from a theoretical perspective and through extensive testing. For example, Buchholtz et al. (2023) come to rather negative conclusions on this basis and state that the generative AI ChatGPT is not yet suitable for use in mathematics classes.

While AI has been studied intensively in educational research from a technical perspective for around 10 years (see, e.g. "International Journal of Artificial Intelligence in Education"), mathematics education research has only recently begun to address the topic, as can be seen above. In particular, so-called large language models (LLMs) such as ChatGPT are being considered. LLMs are linguistic models that have been trained with a huge amount of text data and are intended to simulate communication. With the help of probability trees, answers to user requests (so-called prompts) are generated. Although the system was trained for linguistic knowledge, it can also contain rational knowledge from the training data (Petroni et al., 2019). However, knowledge databases are not accessed for the answers; the "knowledge" comes solely from the trained linguistic model, which can also result in the output of incorrect information. Kasneci et al. (2023) explain in the context of LLMs in education:

> "Large language models can help teachers to identify areas where students are struggling, which adds to more accurate assessments of student learning development and challenges. Targeted instruction provided by the models can be used to help students excel and to provide opportunities for further development." (p. 3)

Initial studies on formative assessment with LLMs have already been carried out. Moore et al. (2022) used a fine-tuned GPT-3 model to evaluate student answers in chemistry education and concluded that it is a powerful tool to assist teachers in the quality of their evaluations of students. Zhu and Liu (2020) found that LLMs can support high school students in scientific reasoning in the area of climate activity. Sailer et al. (2023) observed in a teacher education program that the use of LLMs leads to better justifications of diagnoses of students' learning difficulties.

In this article, the opportunities and challenges of ChatGPT as a formative assessment and feedback tool in mathematics education will be discussed. For this purpose, the LLM is tested and analyzed in detail against the background of a previously described framework based on Fahlgren et al. (2021).

Digital formative assessment and feedback – A framework for the analysis

The basis for the analysis framework in this article is the survey report by Fahlgren et al. (2021) on technology-rich assessment in mathematics. In this report, the research or development projects STACK, STEP and SMART are analyzed and compared against the background of selected categories. The categories considered include localization, receiver, content and form, and adaptivity. Localization refers to whether the assessment and feedback takes place on a micro-level (e.g. task level) or a macro-level (e.g. overall performance). The receiver of feedback is the person to whom the feedback is directed (e.g. student, teacher). The content of feedback can also differ (e.g. right/wrong, hints or error information, worked out examples) as well as the form in which it is presented (e.g. language, pictures). Adaptivity refers to the extent to which aspects of the student response appear in the feedback.

For the analysis of ChatGPT as a potential feedback and assessment tool in this article, the above categories were slightly modified. For this purpose, the aspect of localization was expanded to include the possibilities of input by the user. The adaptivity and receiver categories were combined into one analysis aspect. The aspect of assessment quality, which is important to examine in the field of generative AI, has been added. The analysis in this article is therefore based on four aspects:

- Input and localization
- Assessment-quality
- Content and form of feedback
- Adaptivity and receiver of feedback

Evaluation of the large language model ChatGPT

The LLM analyzed in this article is the latest version of ChatGPT 4.0 at the time of the analysis. Extensive testing was carried out with this LLM in January 2024. The Wolfram plugin, which establishes a connection to the Wolfram Alpha computer algebra system (CAS), was used for all user requests. The testing was based on selected tasks in the field of linear algebra and analytical geometry from the publicly accessible mathematics secondary school examinations in North Rhine-Westphalia (Germany) (see https://www.standardsicherung.schulministerium.nrw.de/cms/zentralabitur-gost/faecher/fach.php?fach=2). The tasks include, for example, an inner-mathematical task in which distances between two points had to be determined and a third point had to be chosen so that a right angle is formed, or an application task with a real-world context in which it had to be justified that the base of a pyramid lies in one of the coordinate planes, it had to be shown that three given corner points of the pyramid approximately form an equilateral triangle with a certain edge length, and the plane in which these three points are located had to be determined. The tasks were translated into English and formulas were converted into LaTex notation. In addition, sample solutions with different types of errors were created, which were then entered into ChatGPT with different prompts. The conversations with ChatGPT formed the data basis for the analysis.

The results of the analysis with regard to the four aspects mentioned above are presented below and explained at selected points with examples from the testing. For readability, the explication is limited to one task in which a linear system of equations is to be solved. It should be emphasized at this point that although the analysis framework used was precise and comparable with other studies, the data collection was rather exploratory and not systematic, which is why this is not a scientific study, but rather an experience report.

Aspect 1: Input and localization

As described above, LLMs like ChatGPT are developed especially for processing text and are intended to simulate communication. Therefore, the input and also the output is mainly symbolic as text. If the Wolfram plugin is switched off, it is also possible to input images or sound recordings, although the analysis options in this case are very limited beside the recognition of text on the images or in the recordings.

Mathematical formulas can be entered in any programming language or as a kind of pseudo-code. In this experiment, formulas were entered as LaTex codes. Various external software is available that automatically converts handwritten formulas or texts into equivalent LaTex codes, which can then be copied into the text input field of ChatGPT.

To enable assessment and feedback by ChatGPT, both the task and the user's own solution must be entered. To complete the prompt, it is also necessary to explain what is to be done in relation to the task and the solution (e.g. "Can you tell me if this is correct?"). Characteristic for an LLM such as ChatGPT is the possibility to ask follow-up questions after the response of the system, which then also changes the feedback (e.g. "Please tell me where exactly the error occurs."). The further conversation automatically includes the previous requests and answers.

However, retaining the context of a conversation is only possible to a certain extent. The number of tokens that can be used (8192 tokens for ChatGPT 4.0 at the time of analysis) limits the number of analysis units included, whereby according to ChatGPT, one token corresponds to approximately four characters. This means that the localization of the assessment and feedback provided by ChatGPT is more likely to be at the micro level. Although several tasks and solutions can be included, a structured modeling of learning paths over a longer period of time is not possible.

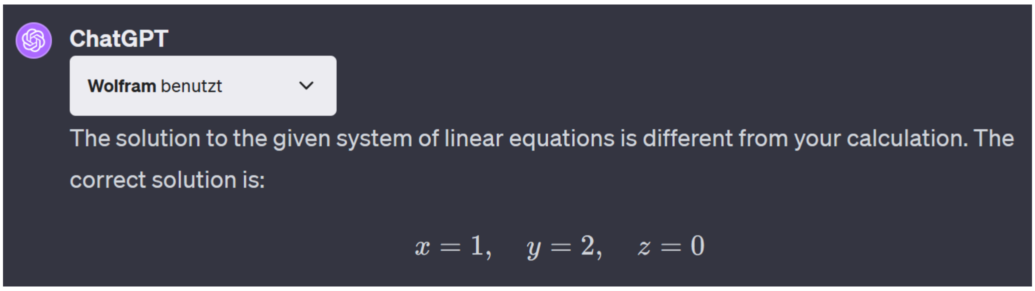
Aspect 2: Assessment-quality

In terms of mathematical correctness, the detailed testing revealed that many of the calculated results were correct and that the feedback on the user's solution was also correct on this basis. The Wolfram plugin establishes the connection between the LLM and the CAS Wolfram Alpha. This means that potentially all operations that can be performed by a CAS (e.g. transformation of equations, calculation of derivatives and integrals) can also be used by ChatGPT. Nevertheless, ChatGPT may misinterprets information when analyzing the task and thus sends an incorrect request to Wolfram. It is also possible that ChatGPT misinterprets the results from Wolfram and thus provides incorrect feedback to the user. However, in the responses from ChatGPT it is made transparent which requests were sent to Wolfram and which responses were given. This means that errors can be quickly identified in most cases. Errors also frequently occur if the Wolfram plugin is not used. This can be problematic, as the reasoning around the incorrect calculations can still be plausible and therefore potentially not recognized as wrong by non-experts (Buchholtz et al., 2023). It is therefore recommended to always explicitly state in a request that the Wolfram plugin should be used.

In general, ChatGPT has proven to be quite reliable in processing math problems, even when real-world contexts occur. However, there are actually no reliable figures on mathematical correctness of ChatGPT responses. Some uncertainties occur in the answers of ChatGPT when conceptual questions are asked (e.g. "What is a probability?"). Problems also arise when mathematical processes outside the capabilities of a CAS are requested. For example, ChatGPT is comparatively unreliable at outputting mathematical proofs or evaluating given proofs. This is mainly due to the fact that LLMs are not mathematically logical systems. Although proofs of classical mathematical theorems can be generated, circular reasoning or incorrect derivations often occur.

Aspect 3: Content and form of feedback

The feedback that ChatGPT provided on the user's solutions in the testing had different contents. If no further information on the desired content is provided (e.g. only asking "Can you tell me, if my solution is correct?"), ChatGPT usually presents the correct solution and uses this as the basis for judging whether the user's solution is correct or incorrect. Figure 1 shows an example in which a system of linear equations consisting of the equations $4x + 2y - 3z = 8$, $2x - 3z = 2$ and $6x + 2y - 5z = 10$ had to be solved. The system of equations was solved by hand on paper, whereby a transformation error was deliberately included in the first step. The entire solution was converted into a LaTex code and entered together with the question "I have calculated like this. Can you tell me if this is correct? Use the Wolfram Plugin for your calculations.". The response from ChatGPT can be seen in Figure 1. The Wolfram plugin was used to calculate the solution, and this was displayed in the response. In addition, feedback was given that the calculated solution differs from the user's input.

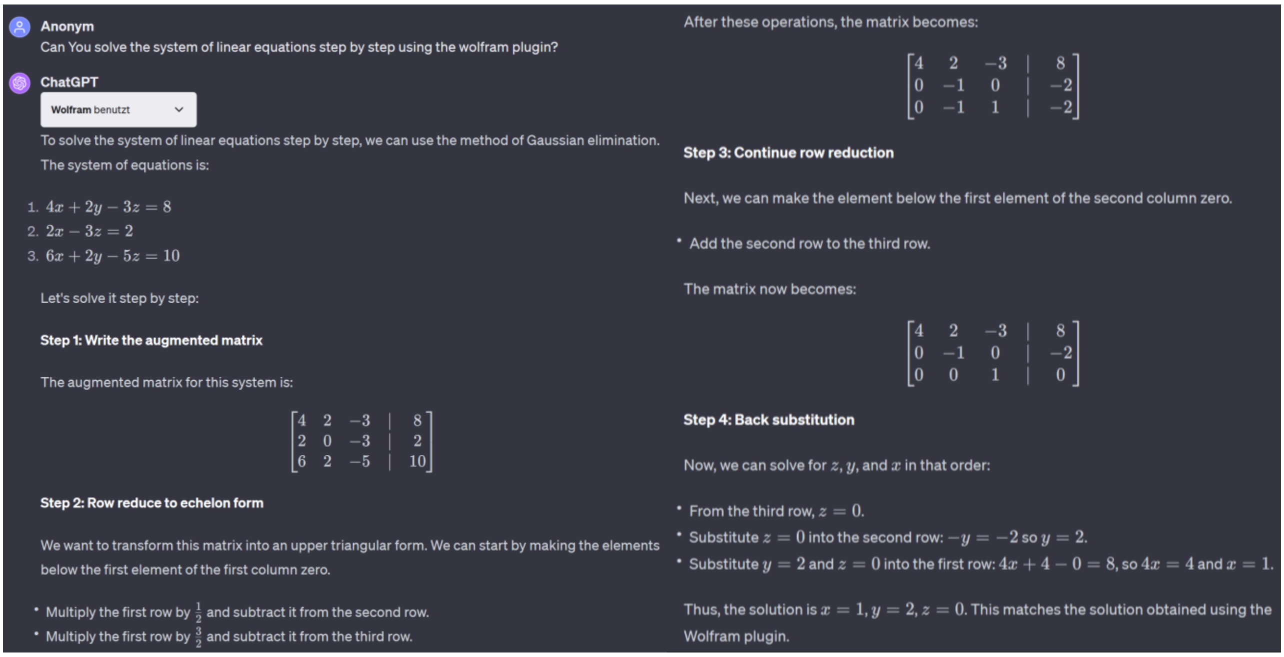


The solution to the given system of linear equations is different from your calculation. The correct solution is:

$$x = 1, \quad y = 2, \quad z = 0$$

Figure 1: Right-Wrong feedback by ChatGPT

When determining the correct solution, the solution process is often presented directly as a kind of sample solution and provided with some explanations. However, this feedback can also be deliberately generated using suitable prompts, e.g. "Can you solve the task step by step and explain it?". Figure 2 shows such a query in relation to the task above. ChatGPT first writes down the system of linear equations to be solved and explains that this can be solved using the method of Gaussian elimination. The subsequent solution process is divided into four steps. In the first step, the augmented matrix is written down. The second and third steps consist of transformations that convert the matrix into a triangular form. In the fourth step, the concrete values for x, y and z are determined by substitution. The solution is described in relatively great detail and some instructions are given on why the individual steps should be carried out.



Anonym

Can You solve the system of linear equations step by step using the wolfram plugin?

ChatGPT

To solve the system of linear equations step by step, we can use the method of Gaussian elimination. The system of equations is:

1. $4x + 2y - 3z = 8$
2. $2x - 3z = 2$
3. $6x + 2y - 5z = 10$

Let's solve it step by step:

Step 1: Write the augmented matrix

The augmented matrix for this system is:

$$\left[\begin{array}{ccc|c} 4 & 2 & -3 & 8 \\ 2 & 0 & -3 & 2 \\ 6 & 2 & -5 & 10 \end{array}\right]$$

Step 2: Row reduce to echelon form

We want to transform this matrix into an upper triangular form. We can start by making the elements below the first element of the first column zero.

- Multiply the first row by $\frac{1}{2}$ and subtract it from the second row.
- Multiply the first row by $\frac{3}{2}$ and subtract it from the third row.

After these operations, the matrix becomes:

$$\left[\begin{array}{ccc|c} 4 & 2 & -3 & 8 \\ 0 & -1 & 0 & -2 \\ 0 & -1 & 1 & -2 \end{array}\right]$$

Step 3: Continue row reduction

Next, we can make the element below the first element of the second column zero.

- Add the second row to the third row.

The matrix now becomes:

$$\left[\begin{array}{ccc|c} 4 & 2 & -3 & 8 \\ 0 & -1 & 0 & -2 \\ 0 & 0 & 1 & 0 \end{array}\right]$$

Step 4: Back substitution

Now, we can solve for z, y, and x in that order:

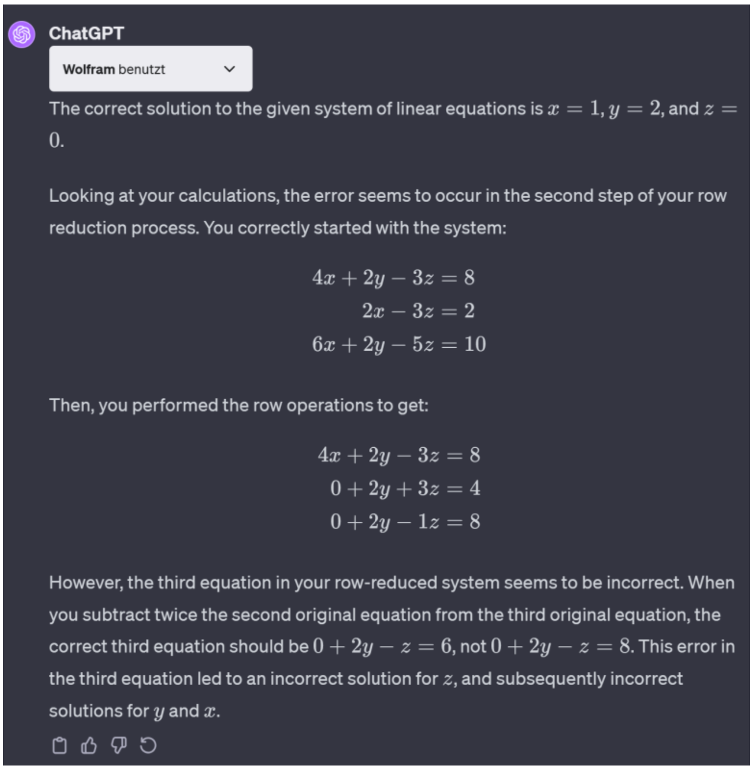
- From the third row, $z = 0$.
- Substitute $z = 0$ into the second row: $-y = -2$ so $y = 2$.
- Substitute $y = 2$ and $z = 0$ into the first row: $4x + 4 - 0 = 8$, so $4x = 4$ and $x = 1$.

Thus, the solution is $x = 1, y = 2, z = 0$. This matches the solution obtained using the Wolfram plugin.

Figure 2: Solving and explaining a task step by step by ChatGPT

In most cases, an explicit prompt is required for a specific feedback content. For example, it is possible to ask at which point the error occurs. ChatGPT can specify this sometimes more and sometimes less precisely and correctly. One of the reasons for this is that the Wolfram plugin does not output the solution path to ChatGPT, but only the final results – against this background, the solution path must therefore be checked by ChatGPT itself.

In Figure 3, for the example above, the user asks where exactly the error is located in the calculation. ChatGPT then checks whether the row operations from the first step are correct and determines that an incorrect equation has been set up in the third line. This is actually the row in which the error was deliberately inserted for the testing. However, when determining the correct equation, ChatGPT makes a mistake, too, and states that the equation $0 + 2y - z = 6$ is created by double subtracting the second line from the third line. This operation actually results in $2x + 2y + 1 = 6$, which does not help to solve the system of equations.



The correct solution to the given system of linear equations is $x = 1, y = 2$, and $z = 0$.

Looking at your calculations, the error seems to occur in the second step of your row reduction process. You correctly started with the system:

$$\begin{aligned} 4x + 2y - 3z &= 8 \\ 2x - 3z &= 2 \\ 6x + 2y - 5z &= 10 \end{aligned}$$

Then, you performed the row operations to get:

$$\begin{aligned} 4x + 2y - 3z &= 8 \\ 0 + 2y + 3z &= 4 \\ 0 + 2y - 1z &= 8 \end{aligned}$$

However, the third equation in your row-reduced system seems to be incorrect. When you subtract twice the second original equation from the third original equation, the correct third equation should be $0 + 2y - z = 6$, not $0 + 2y - z = 8$. This error in the third equation led to an incorrect solution for z, and subsequently incorrect solutions for y and x.

Figure 3: Locating the error by ChatGPT

Many other feedback contents can be generated using suitable prompts. For example, a mathematical process that is related to the task can be explained again. In relation to the task above, for example, you can ask how linear systems of equations can be solved. In a test, ChatGPT has listed different methods (graphical, substitution, elimination, Gaussian matrix elimination, Cramer's rule, software or calculators), briefly explained them and compared them according to the criteria "applicable for" and "best for".

In addition, further exercises including sample solutions can be generated. For this purpose, certain criteria can also be specified for the task, e.g. that only integers should occur in the matrix or that the task is integrated into a real context. A query for the above task resulted in an application, that concerns the relationship between costs and the manufacture of three different products in a company. This is certainly an authentic context in the field of linear equations. However, the task resulted in a system of linear inequalities. From a didactical perspective, it is questionable whether this is a good subsequent task for students struggling with solving systems of linear equations.

Finally, it should be emphasized at this point that it is also possible to deliberately avoid certain feedback content in ChatGPT. This is particularly interesting with regard to the output of the solution. For example, prompts can be formulated in which it is specified that the solution should not be given to the users under any circumstances, but only hints should be provided to help them solve the task or find the errors themselves. A customized version of ChatGPT called "Soctratic Tutor" is already provided specifically for this purpose. If the user asks the tutor how to solve the system of equations mentioned above, the tutor will ask suitable questions in order to help the user actively work on the task step by step (e.g. “Can you identify how many variables and how many equations are presented in this system? And why is this information important for solving a system of linear equations?”).

In addition to the content of the feedback, ChatGPT also allows different forms of presentation. The focus is on feedback in written form. ChatGPT uses in particular mathematical terminology, which is characterized by the corresponding mathematical terms. However, isolated elements of colloquial language can also be identified in the conversation. In relation to the tasks, algebraic expressions can be found at many points. While the input by the user is carried out using LaTex codes or similar codes, which are somewhat confusingly displayed, the formulas in the ChatGPT responses are easy to read (Figures 2-3). In addition to language feedback, ChatGPT can also output iconic representations. For plotting the above linear equation system, for example, it can use the Wolfram plugin and display an appropriate three-dimensional graphic of three intersecting planes.

Aspect 4: Adaptivity and receiver of feedback

The intensive testing has shown that the level of adaptivity of ChatGPT can be very high. For example, the feedback always refers specifically to the task set at the beginning. Instead of giving a general description of the solution procedure, the solution path can be presented step by step according to the task. When checking the user's own solution, ChatGPT can even localize errors in the solution path to a certain extent and is not limited to comparing the final results with the results determined by the Wolfram plugin.

In addition to adaptivity in relation to the task and the solution to be checked, suitable prompts can also be used to make adjustments in relation to the user's characteristics. For example, a studied mathematician should receive different feedback than a student in secondary school. At the beginning of a chat or in the settings of the ChatGPT account, relevant information can be entered. The more detailed the information is, the more accurately ChatGPT can take it into account in the conversation. For example, language difficulties of the user can be pointed out so that responses use simple words and short sentences. It can also be emphasized as positive that users can work with ChatGPT in the language in which they feel most confident regarding mathematics.

Conclusion

The previous analysis has shown that ChatGPT already offers a remarkable amount of potential for formative assessment and feedback in the field of mathematics. This is in line with the results already obtained in studies outside mathematics (Moore et al., 2022; Zhu & Liu, 2020; Sailer et al., 2023). The right prompting has proven to be an important success factor for appropriate feedback. The prompt largely determines the form and content of the feedback, the correctness of the performed calculations and the extent to which the response is adapted to the feedback receiver.

However, a number of challenges remain. Probably the most important one concerns the mathematical correctness of the calculations performed by ChatGPT, which is not guaranteed. This is particularly problematic because learners often do not have the competencies to recognize the errors. Students therefore need well-developed reflection skills, a critical handling of the system and close support from the teacher.

With regard to the analysis categories according to Fahlgren et al. (2021), the strengths of ChatGPT lie in particular in the adaptivity of the responses in relation to the task and the feedback receiver (Aspect 4) as well as the content and form of the feedback (Aspect 3). The possibilities for input are largely restricted to text and localization is limited to the micro level (Aspect 1). The assessment quality is high for numerical and symbolic calculations as considered in this analysis, but there is generally no certainty for the correctness of the responses (Aspect 2). Thus, LLMs such as ChatGPT cannot replace the assessment and feedback systems developed specifically for learning mathematics – but if used appropriately, they can be suitable additions. The future will show how the development of generative AI will progress and how this will affect the opportunities and challenges identified above.

References

Buchholtz, N., Baumanns, L., Huget, J., Peters, F., Pohl, M., & Schorcht, S. (2023). Herausforderungen und Entwicklungsmöglichkeiten für die Mathematikdidaktik durch generative KI-Sprachmodelle. *Mitteilungen der GDM, 114*, 19–24.

Fahlgren, M., Brunström, M., Dilling, F., Kristinsdóttir, B., Pinkernell, G. & Weigand, H.-G. (2021). Technology-rich assessment in mathematics. In: A. Clark-Wilson, A. Donevska-Todorova, E. Faggiano, J. Trgalova, & H.-G. Weigand (eds.), *Mathematics Education in the Digital Age* (pp. 69–83). Routledge.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences, 103*, 102274.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE. *PLoS digital health, 2*(2), e0000198.

Moore, S., Nguyen, H.A., Bier, N., Domadia, T., & Stamper, J. (2022). Assessing the quality of student-generated short answer questions using GPT-3. *Proceedings of EC-TEL 2022*, 243–257.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases? *arXiv preprint* arXiv:1909.01066.

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction, 83*, 101620

Wardat, Y., Tashtoush, M., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *EURASIA Journal of Mathematics, Science and Technology Education, 19*(7), em2286.

Zhu, O.M., & Liu, H.L. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education, 143*, 103668.

From automatic diagnosis to lesson plannings: how teachers use elements of a digital formative assessment tool

Eumann, Anica; Klingbeil, Katrin and Barzel, Bärbel

University of Duisburg-Essen, Germany; anica.eumann@uni-due.de

As the effectiveness of formative assessment depends on the form of implementation in school lessons, we investigate in this study which elements of the digital formative assessment tool SMART teachers use for which purposes. In an exemplary qualitative analysis of two teacher interviews, we found that teachers use the given teaching suggestions and materials for designing their upcoming lessons. Furthermore, they use the test items and didactical information about (mis)conceptions to professionalize themselves by gaining deep insights into students' thinking.

Keywords: formative assessment, digital tool, algebra lessons, teacher professionalization

Introduction

Empirical studies showed that formative assessment (FA) may have positive effects on students' learning, depending on the subject and the concrete form of implementation (McLaughlin & Yan, 2017). For this reason, Schütze et al. (2018) state a high need for research to find out more about different types of realization of FA in classroom practices.

In this paper, we present first results of a study examining the concrete use of a digital FA tool by teachers in algebra lessons in secondary schools in Germany. The tool used in this study is SMART (Specific Mathematics Assessments that Reveal Thinking) that results from a project of the University of Melbourne and is currently being adapted to German speaking countries. This tool aims at giving precise diagnosis of students' thinking and possible misconceptions for specific topics.

Theoretical Background

FA is defined as an activity in which "evidence about student achievement is elicited, interpreted, and used by teachers, learners or their peers, to make decisions about the next steps in instruction". (Black & Wiliam, 2009, p. 9)

In this process, the teacher as a decision maker plays a very important role. In their model of technology enhanced FA, Cusi et al. (in print) differentiate four areas of teachers' practice in FA (sharing goals and criteria, designing and implementing learning activities, fostering the quality of feedback, involving students in peer- and self-assessment) which they combine with the three phases of preparing lessons (pre-paration, paration and meta-paration) as well as with three functionalities of technology in the process of FA (communicating, analyzing and adapting).

The digital tool SMART that is used in this study consists of about 130 multiple-choice-tests in several mathematical topics that elicit students' understanding of mathematical issues by focusing on their conceptual knowledge. Thus, it allows communication through and with technology because information is displayed and submitted but the user may also interact with the elements of the tool. The tool carries out an advanced analysis on the basis of response pattern that is shown in individual stages of understanding and possible misconceptions displayed in the automatic evaluation for the teacher. Hereby, it allows teachers to get an insight into students' thinking. Adaptation in this tool rests passive because it proposes teaching suggestions and materials for further instruction to the teacher. At least, they decide whether and how they use this supply (Price et al., 2013).

Research Question & Methods

As we have seen, working with SMART demands to a high extend teacher activities. This leads to our research question: How and wherefore do teachers use the components of SMART after the implementation of a test? In this paper, we will focus on the second aspect.

To answer this question, we interviewed teachers after having used the tool in their lessons for the first time. These interviews are part of the project SMART[alpha] in which we investigate in a control-group-design the thinking of students and the development of FA competencies of teachers while working with the tool. In this interview, we followed a guideline consisting of two main issues of discussion: the teachers' impression of the test results and the consequences they draw from it for the upcoming lessons.

In the following, we present first results to this research question, arising from an exemplary analysis of two teacher interviews. These are male teachers that teach in North-Rhine-Westphalia in Germany at two different secondary schools in grades 8 resp. 7. Both teachers were picked from a group that did not participate in a professional development program. The only hints for use provided were on a technical level in the handling of the tool. Thus, they worked with the SMART tool independently and autonomously in their classes and used the automatic diagnosis in their sole discretion for their further lesson plannings.

The data analysis is carried out by a qualitative content analysis following Kuckartz (2018). Hereby, we developed a deductive-inductive category system, in which the deductive categories arise from the four areas of FA practice as well as the three phases of preparing lessons (see Cusi et al., in print).

Results

The data analysis led to three essential aspects of using the tool which will be explained and illustrated by examples.

The first aspect, the *lesson design*, is realized on three different levels that are part of the three deductive categories arising from the phases of preparing lessons. On the level of *pre-paration*, Teacher 1 explains that in the future he will treat typical errors and misconceptions more explicit. Teacher 2 reports that he will use the given material much earlier in future lessons. On the level of *paration*, Teacher 1 reports on the one hand that he uses particular teaching material unchanged and on the other hand that he adapted information on misconceptions and some questions of the test items to discuss them in the classroom. Here, he focusses primarily on the results of the whole class. In addition, teacher 1 also develops new ideas for teaching as for example to work out strategies and rules with the students that might help to avoid typical errors. In contrast, teacher 2 reports that he used the material only in the form in which it is contained in the tool. An adaption or development of new ideas does not take place. Apart from these results, the interviews with both teachers also show that the tool encourages them to reflect their own teaching, which is part of the level of *meta-paration*: "well, as I now knew a bit what is in the tool, and payed a bit more attention to it, also in discussions in class, I noticed more." (Teacher 1).

This statement also indicates the second aspect of use, the *professionalization*. This is shown on two levels arising from two inductively developed categories: first the level of reflection of one's own thinking and practice and second on the level of an intentional knowledge acquisition. Both teachers describe trying to find explanations for particular answers in individual students' results, which is a form of *reflection on their own thinking and practice*. Teacher 2 explicitly names that he is aware of these reflection processes: "Also, ehm, I felt catched, to be honest, so that means, objectively spoken, I reflected obviously." In a similar way, Teacher 1 describes a non-intended professionalization as an "aha-effect" concerning his own language practice in classroom. He formulates the consequence that in future lessons he will pay more attention in the class to being a role model in the use of language and to pay more attention to the students' use of language as well.

Teacher 2 also shows a process of reflection in his work with the teaching suggestions and the information about the levels of understanding and misconceptions. He reports that first he was a little bit annoyed by the length of the texts, but then admits: "I would also say afterward, that I have taken some time to look at it, I think it's great that there is some didactical background. Ehm, I have to admit, that I also took a lot with it." This citation also shows the professionalization on the second level of *intentional knowledge acquisition*. Teacher 1 also mentions this very specifically in the report that he initially had very few ideas about misconceptions, but in the end, recognized them in his own lessons. The most intensely the aspect of knowledge acquisition becomes obvious in the end of the interview with Teacher 2 when he states that the work with the tool was very helpful for him on different levels and that he would use the tool once again but with constraints. The decision to work with the tool depends for him on two factors: the subjective relevance as well as his own pedagogical content knowledge of the concrete topic. This leads to the conclusion that he actively and intentionally uses the tool to gain new knowledge.

The third aspect, the *exploration of the automatic diagnosis,* also arises from inductively developed categories. On the one hand, teacher 1 tries to retrace the automatic evaluation by trying to find connections between different response options in the items and the information about the stages of understanding and misconceptions. On the other hand, he notices that some students do not show the results he would have expected so that he develops explanations that take into account students' thinking: "Well, I found that very fascinating. […] that you can somehow understand what could be the way of thinking that a student had." Teacher 2 also recognizes a high discrepancy between the results of the automatic diagnosis and his own observations in classrooms. This is why he looks at some individual answers given by students to comprehend the automatic analysis. He looks deeply into the content of the items and the didactic information which allows him to understand that it is not the number of correct answers that leads to a certain stage of understanding but the type of answers given. Thus, he also gains deep insights into students' thinking that he tries to put into relation with his own practice in classroom.

All in all, we find that both teachers put into effect similar activities while working with the tool that differ clearly in their concrete specification. While Teacher 1 focuses on the teaching suggestions and materials to design upcoming lessons by planning new impulses and developing new ideas out of the given materials, Teacher 2 concentrates on the test items and the didactical information to understand the automatic diagnosis, reflect his own practices and educate himself on a didactical level. But in both cases, we can see an active change of lesson design and a process of teacher professionalization.

Discussion & Outlook

The results give a hint that the teaching materials are actively used in mathematics lessons but that we have to differentiate between an unmodified and an adapted use. Moreover, the case of Teacher 1 shows that the test items as well as the didactical information about students' (mis)conceptions are used to plan and reflect lesson practices. Particularly, the information is able to encourage teachers to develop their own new ideas of lesson activities. We have also seen that the test items as well as the teaching suggestions and materials support teachers in gaining new pedagogical content knowledge and reflecting their own thinking and practice so that the work with the elements of the tool may contribute to teacher professionalization.

Didactically deep founded diagnostic items as well as didactical background information and fitting teaching suggestions and material therefore seem to be important elements of digital FA tools that teachers use intensively. It is to be underlined that the passive adaptation of SMART is sufficient for such an intensive use.

The presented study nevertheless underlies some restrictions. At the moment of the interviews, the teachers temporarily had a restricted access to the results of the diagnosis so that they had to reconstruct some aspects based on their memory. Moreover, these are all only self-reported practices where it is impossible to check their real implementation. But this also offers a new research perspective to a project in which teachers will be attended in their work with SMART and the use and implementation of the tool will be examined by lesson videography.

References

Cusi, A., Aldon, G., Barzel, B. & Olsher, S. (in print). Rethinking teachers' formative assessment practices within technology-enhanced classrooms. In B. Pepin, G. Gueudet & J. Choppin (Eds.), *Handbook of Digital Resources in Mathematics Education, Ed 1*.

Black, P. & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Kuckartz, Udo (2018). Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung. Beltz Juventa.

McLaughlin, T. & Yan, Z. (2017). Diverse delivery methods and strong psychological benefits. A review of online formative assessment. *Journal for Computer Assisted Learning, 33*, 562–574.

Price, B.; Stacey, K.; Steinle, V. & Gvozdenko, E. (2013). SMART ONLINE ASSESSMENTS FOR TEACHING MATHEMATICS. *Mathematics Teaching, 235*, 10–15.

Schütze, B.; Souvignier, E. & Hasselhorn, M. (2018). Stichwort – Formatives Assessment. *Zeitschrift für Erziehungswissenschaften, 21*(4), 697–715. https://doi.org/10.1007/s11618-018-0838-7

Peer assessment in an undergraduate geometry course: Fostering proof competency in teacher students

Yael Fleischmann[1], Antoine Julien[2] and Alexander Schmeding[1]

[1]Norwegian University of Science and Technology, Trondheim, Norway; yael.fleischmann@ntnu.no, alexander.schmeding@ntnu.no

[2]Nord universitet i Levanger, Norway; antoine.julien@nord.no

Proofs are integral to mathematics as a science, but they are difficult to learn and provide challenges for instruction. This is a particular problem in teacher education, where proofs are a topic that many students experience being disconnected from teachers' day-to-day work in schools. In this note, we report on a course development and research project exploring the use of peer assessment as a tool to foster proof competency in pre-service teacher students.

Keywords: Assessment, teacher education, mathematics education, proof competency, mathematical proofs, peer assessment

Introduction

Proofs are both a defining product of mathematicians' activity and a notoriously difficult topic for undergraduate students of mathematics. The usual mathematics curriculum in schools puts little emphasis on proofs, arguments, and formal explanations which is a source of tension for the design of courses for pre-service teachers (PSTs). On the one hand, students need to acquire competency in producing and understanding proofs and arguments; on the other hand, the level of formality of tertiary education proofs is often viewed by PSTs as irrelevant to the practice of teaching in schools.

In this article, we present a course development project aimed at training PST's proof competency and changing their attitudes on the topic. Our method to attain these goals is to use peer assessment. The study detailed in this paper is set to commence in early 2024. Note that this paper therefore will not feature results or data from this ongoing study. We anticipate sharing preliminary findings at the FAME conference. The design for the study draws inspiration from an earlier study which we conducted with the aim to improve computational skills of PSTs through peer assessment. Our overarching aim is to develop peer assessment as a tool to enhance *proof competency* (refer to the theory section for the definition of this term used throughout the paper) of the PSTs. The present article more modestly restricts to the following research questions:

1. To what extent does peer assessment contribute to the improvement of proof understanding and proof construction skills among PSTs participating in an undergraduate geometry course?
2. How effectively can PSTs assess the clarity and logical soundness of proofs and arguments generated by their peers during a peer assessment activity?

To this end, we incorporate peer assessment into a geometry course for teacher students. In the implemented activity, the PSTs are asked to evaluate each other's mathematical argumentation. Our interest here lies in the effects of peer assessment on proof understanding and construction abilities of the PSTs. Note that the evaluation of mathematical arguments is typically an unfamiliar task for PSTs in Norway.

Theory and background

Assessment and feedback can be effective tools teachers can use to promote students' learning (Hattie, 2008). For this study we are interested in peer assessment as a tool. Topping (1998, p. 250) defines peer assessment as being an arrangement in which students evaluate the work of peers of similar status. In a previous study (Julien, Romijn, Schmeding 2023) we investigated how peer assessment, in particular giving and receiving of feedback, enhanced mathematical knowledge of PSTs. There, assessment activities which involved giving feedback were shown to have potential to both enhance students computational and professional skills. These two objectives can be related to Shulman's distinction between pedagogical content knowledge (PCK) and subject matter knowledge (SMK), see e.g., Berry et al. (2016). We focus on assessment in mathematical tasks which address SMK as in mathematical competency but not PCK. In the literature, the effect of peer evaluation on PCK for pre-service mathematics teachers is discussed in Ayalon & Wilkie (2021). In contrast, there is little research on the influence of peer assessment on SMK in mathematics teacher education.

The purpose of the present paper and the associated study is to investigate the effects of peer assessment on proof understanding and construction in an undergraduate geometry course. As Lin et al. (2012) stress, teachers professional learning of proofs and teaching proofs depends on their knowledge, practice, and beliefs about proofs. Our main goal is to investigate how peer assessment among PSTs can be used to develop those aspects.

So far in this article, the term "proof competency" was used as an umbrella term which has not yet been defined; we shall remedy it now. The teaching of mathematical arguments and proofs in higher education and in teacher education has been an active research subject for quite some time. With a view towards teacher education, there are three different aspects to be considered: knowledge of proof, practice of proof and beliefs about proof. These are interdependent and need to be addressed simultaneously to improve proof competency (Lin et al., 2012). For our study, the construct of proof competency consists of four related aspects (see Selden and Selden, 2015): proof comprehension, proof construction, proof validation and proof evaluation.

Proof comprehension is the ability to read and understand written proofs. The "big difference" (Selden and Selden 2015, p. 4) between proof comprehension and proof validation is that in a proof comprehension situation, it can be assumed that the presented proof is correct, while this is not the case in proof validation situations. The distinction is of particular importance for us, as PSTs train for situations in which they are asked to validate and evaluate arguments and proofs. Proof evaluation describes the assignment of a value judgement to a proof (attempt). For professional mathematicians this often means judging a proof on its merits of conveying ideas and concepts. We view it as equally important for PSTs to be able to assess the presentation and clarity of a proof.

The three aspects of proof competency described in the last paragraph have in common that they apply to proofs presented to the PSTs. In contrast, proof construction asks for the creation of new proofs, usually to a statement provided to the learners. This activity in general is inherently more challenging than the other aspects and often requires substantial SMK. However, in a sense also the PCK is called upon in the construction and presentation of arguments to convince the reader of the validity of a claim. This social dimension can be viewed through the lens of communities of practice (here the PSTs in the course), see e.g. Selden (2012, section 3.1.2). Proofs need to provide an acceptable level of conviction that the mathematical statement is true. Following Mason et al. (1982) these levels are (in ascending order of sophistication): 1. Convince oneself, 2. Convince a friend, 3. Convince a sceptic. What is viewed as sufficient to qualify for the different stages in the model will depend on the social norms and practices within the community for which the proof is constructed.

Methodology and setting

The setting in which we will carry out the peer assessment is a geometry course in a large Norwegian university, which has a particular focus on axiomatic constructions. One of the main learning goals of this course is to revisit classic geometric results from a higher standpoint. This includes explicit proofs in an axiomatic setting. Thus, the course aims at developing students' geometric understanding as well as their understanding of and ability to produce proofs. The main stated public for this course is pre-service high-school teachers in their second or third study year.

In one of the assignments which the students need to hand in, they will assess an educator-made proof (based on student deliveries from a previous iteration of the course). This preparation task was selected to display subtle aspects of proving, which students usually find difficult, such as the need to prove that a condition is both necessary and sufficient. Two actual "peer assessment events" will then be carried out during the semester. Such an event consists of the following: students will solve a task knowing that it will be assessed by another student, and hand it in. In the next homework assignment, they will receive one of their peers' solutions from the previous assignment and assess it. For this, the PSTs will be given a grading guide. Our design aim for the guide is to strike a balance between general and specific instructions: the guide needs to be specific to provide scaffolding for the assessment; nevertheless, it should also not be a step-by-step solution as we want students to exercise their own judgment and autonomy in assessing statements and justifications.

To answer our research questions, we will evaluate both the quality of the students' proofs and the product of the peer assessment of these proofs. For the second research question, we will focus on clarity and logical soundness of arguments, as reflected in the proof and in the peer assessment of the delivery. We will then compare our own assessments with those obtained by the students in the peer assessment process. In addition, we will conduct interviews with participants. This qualitative data together with the results of a formal written assessment of the proof understanding and construction (conducted three times during the semester) will allow us to create a holistic description of the PSTs abilities. The assessment of proof understanding and construction is also part of a second research project focusing on the influence of learning videos on proving skills. The results from the assessments and additional interviews will shed light on the first research question.

Discussion

In designing our peer assessment experiment and data collection, we needed to clarify our goals: a large aspect of learning proofs is a question of students' autonomy. As Robert and Schwarzenberger point out "tertiary students need to learn to distinguish between mathematical knowledge and meta-mathematical knowledge of the correctness, relevance and elegance of proof and take responsibility for their own mathematical learning." (cited from Guzman et al., 1998, p.755). With that in mind, we believe that students having the responsibility of establishing the correctness of a peer's proof might be of value. In addition, reading the grading guide will make explicit and visible to them the set of demands that mathematicians make on what can be called a “proof”. This relates to the “enculturation” aspect of learning proofs: evaluating other people’s reasoning is an authentic activity both for mathematicians and mathematics teachers.

We are convinced that peer assessment activities can be developed to become a valuable tool for the acquisition of proof competency and the professionalization of teacher students.

References

Ayalon, M., & Wilkie, K. J. (2021). Investigating peer-assessment strategies for mathematics pre-service teacher learning on formative assessment. *Journal of Mathematics Teacher Education, 24*(4), 399–426. https://doi.org/10.1007/s10857-020-09465-1

Berry, A., Depaepe, F., & van Driel, J. (2016). Pedagogical content knowledge in teacher education. In J. Loughran & M. L. Hamilton (Eds.), *International handbook of teacher education: Vol. 1* (pp.347–386). Springer. https://doi.org/10.1007/978-981-10-0366-0_9

De Guzmán, M., Hodgson, B. R., Robert, A., & Villani, V. (1998). Difficulties in the passage from secondary to tertiary education. *In Proceedings of the international Congress of Mathematicians* (Vol. 3, pp. 747–762). Documenta Mathematica. https://doi.org/10.4171/dms/1-3/72

Hattie, J. (2008). *Visible learning*. Routledge.

Julien, A., Romijn, E. & Schmeding, A. (2023). *Peer and self-evaluation as a tool in teacher education*. [Submitted for publication]. http://dx.doi.org/10.13140/RG.2.2.13083.64803

Lin, F.-L., Yang, K.-L., Lo J.-J., Tsamir, P., Tirosh, D. & Stylianides, G. (2012). Teachers’ professional learning of teaching proof and proving. In G. Hanna and M. de Villiers (Eds.), *Proof and Proving in Mathematics Education* (pp. 327–346). Springer. https://doi.org/10.1007/978-94-007-2129-6_14

Mason, J., Burton, L., & Stacey, K. (1982). Thinking mathematically. Addison-Wesley.

Mejia-Ramos, J. P., Fuller, E., Weber, K., Rhoads, K., & Samkoff, A. (2012). An assessment model for proof comprehension in undergraduate mathematics. *Educational Studies in Mathematics, 79*(1), 3–18. https://doi.org/10.1007/s10649-011-9349-7

Selden, A. (2012) Transitions and proof and proving at tertiary level. In G. Hanna and M. de Villiers (Eds.), *Proof and Proving in Mathematics Education* (pp. 391–420). Springer. https://doi.org/10.1007/978-94-007-2129-6_17

Selden, A. & Selden, J. (2015). A comparison of proof comprehension, proof construction, proof validation and proof evaluation, In R. Göller, R. Biehler, R. Hochmuth, and H.-G. Rück (Eds.). *Didactics of Mathematics in Higher Education as a Scientific Discipline – Conference Proceedings* (pp. 339–345). Universitätsbibliothek Kassel. http://www.urn.fi/urn:nbn:de:hebis:34-2016041950121

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276. DOI: 10.3102/00346543068003249

Formative assessment: quick surprise quizzes online in class in mathematics higher education

Sandra Gaspar Martins

Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal; sandra.gaspar.martins@isel.pt

Centro de Investigação em Ciências Sociais, Universidade Nova de Lisboa, Portugal; sicgm@campus.fct.unl.pt

In literature, moments of Active Learning in which all students are working effectively are considered an added value. Even better if, at the same time, they provide a formative assessment opportunity in which students receive immediate feedback. Thus, in a Calculus curricular unit, 31 online surprise quizzes were administered in Moodle to 112 students, distributed across nearly all classes. Students can repeat the quizzes as many times as necessary until they obtain the correct answer, without grade deductions. In an anonymous online survey, almost every student indicated that the quizzes were useful. Although they were not mandatory, the participation rate was very high. Many mentioned they were more attentive due to the quizzes, finding them useful for feedback, for understanding the level they were reaching, and for learning new things. Teachers also confirm that it is a pedagogical strategy worth maintaining.

Keywords: Quizzes, Moodle, formative assessment, active learning, feedback.

Introduction

Quizzes are part of several successful approaches with different kinds of students, both in top universities and in other higher education institutions. Examples of these include TEAL (Dori & Belcher, 2004) at the Massachusetts Institute of Technology; SCALE-UP (Beichner et al., 2007) at North Carolina State University; Peer Teaching (Lasry et al., 2008) at Harvard University; Online Learning Modules (Hill et al., 2015) at the University of Sydney.

Particularly, in the teaching of mathematics in Higher Education, there are many different strategies to apply quizzes: as either formative or summative assessment; online or in-class; mandatory or optional; weekly or with other periodicity; generate new instances for each student or not; give penalties for submitting the answer more than once or not; only multiple-choice questions or more sophisticated ones, etc. Researchers are still looking for the best combination. Some approaches can be found in Siew (2003), Varsavsky (2004), Blanco et al. (2009), Lim et al. (2012), etc.

The National Center for Public Policy and Higher Education in the U.S.A. (Twigg, 2005) refers to computer-based continuous assessment and feedback as a key strategy for quality improvement. Shorter and Young (2011) made a comparison of three assessment methods: (1) daily in-class quizzes, (2) online homework, and (3) project-based learning. They found 'daily in-class quizzes' to be the best predictors of students' learning for 117 undergraduate Calculus students.

Making surprise quizzes carried out during classes makes quizzes an Active Learning activity, that is, a moment in which all students are actively working. This adds even more value since it provides immediate feedback to students (Booth, 2012), forces students to be more attentive in classes and increases competitiveness (Nadeem & Al Falig, 2020).

Taking all this into account, the online surprise quizzes described below were applied in class.

Context

This research took place in the school year of 2023/24 in the course of Mathematics Applied to Engineering (Calculus with applications) belonging to the first year, first semester of the graduation in Computer and Multimedia Engineering of the Polytechnical Institute of Lisbon, Portugal. The researcher has been responsible for this course for some years and this year surprise quizzes were introduced in class. Weekly quizzes have been held for several years, but outside of classes. These quizzes appear from time to time in classes, typically after finishing a subject. In general, there is one quiz per class, sometimes at the end of the class, sometimes in the middle and, sometimes at the beginning – referring to the material that was taught in the previous class. We have three classes per week, lasting 1h30 each and we made a total of 31 quizzes.

Some of these quizzes have different questions for each student (randomly generated using some variables), but most are the same for all students. Everyone can try to answer the quiz as many times as they want, the grade will not be discounted. It is natural for students to talk among themselves, compare the results, and end up all having the full score. This is not discouraged, on the contrary, mutual help is encouraged. These quizzes are not mandatory and count very little towards the final grade. They only count if the grade in the exams is higher than 9.0 and if they improve their grade, in which case, counts 5% of the final grade.

The quizzes were mostly used from the weekly quizzes that had already been created for the course, some were new. Both these quizzes were made available on Moodle -– the Learning Management System of the Institute. These quizzes were simple, we tried to choose or create simple questions that do not use much class time. The quizzes were mostly answered in class, but sometimes there was no time to finish and then the students finished at home. The quizzes were only available to students of the class that was taking that quiz, from the time it opened, until the end of that class (typically).

The course had 112 subscribed students, with 6 being considered ghost students since never answered any class quiz, weekly quiz, test, or exam. Classes are not mandatory. The students who went to any class were 103, distributed this way: the researcher was a teacher in two daytime classes, T11D with 37 students and T12D with 36 students. The other teacher on the course had a night class T11N with 10 students and a daytime class T13D with 20 students.

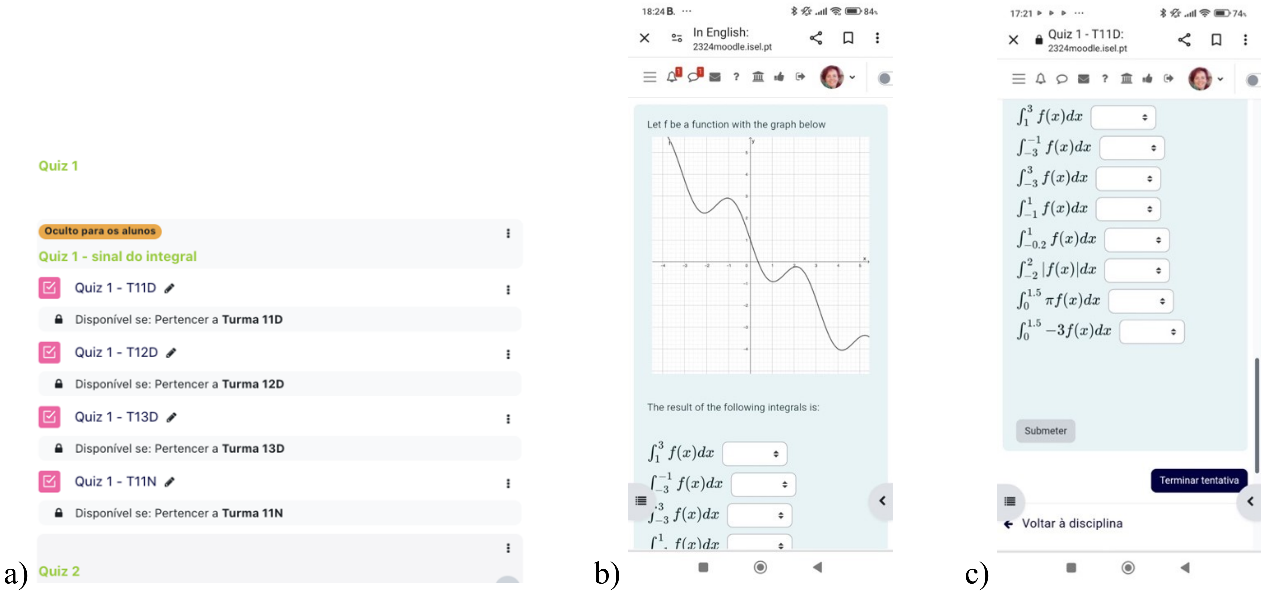
The aim of the quizzes was not to assess students, but to make them be more attentive in class, study more, not to postpone, not to study first the other subjects that naturally are more pleasant to them (since they belong to their study area), and to make students be aware of their level of understanding (often students only realize that they cannot solve the exercises when they get the first test -- in the middle of the semester). Students usually are optimistic about their capabilities (Wandel, 2015) and quizzes help them to be realistic. It was written in Moodle and teachers repeatedly reminded students that the aim of quizzes was to make students be more attentive, study more and be aware of their level of understanding; students may copy all quizzes, but probably will not get the requested values in ‘regular’ assessment and it will not be worthwhile.

One of the advantages of quizzes being online is that they are self-corrected, and we can present them to large classes with little effort to create them and no effort to correct them. Maintaining the advantage of scaffolding the questions as well as in paper and pencil.

There were quizzes on all subjects. Typically, the teacher explained a subject and at the end made a quiz about that subject. It's interesting that in some classes, even in the last few days, when the quizz opened, everyone was attentive and working.

The quizzes

The quizzes were produced through the 'Moodle activity': 'test'. We created one quiz and then use the "duplicate" option to replicate it to all the four classes (Figure 1.a)). However, for the next year, we will create just one quiz and open and close the same test every time we are in a class to create just one quiz. Students don't know what the quiz will be about (it's hidden, only visible to the teacher). Quizzes allows the introduction of images, and mathematical symbols using LaTeX, see Figure 2. Some students answer it on the smartphone others on laptops. It perfectly fits a smartphone, as we can see also in Figure 1.b) and c).



Quiz 1

Oculto para os alunos

Quiz 1 - sinal do integral

Quiz 1 - T11D

Disponível se: Pertencer a **Turma 11D**

Quiz 1 - T12D

Disponível se: Pertencer a **Turma 12D**

Quiz 1 - T13D

Disponível se: Pertencer a **Turma 13D**

Quiz 1 - T11N

Disponível se: Pertencer a **Turma 11N**

Quiz 2

Let f be a function with the graph below

The result of the following integrals is:

$\int_1^3 f(x)dx$

$\int_{-3}^{-1} f(x)dx$

$\int_{-3}^{3} f(x)dx$

$\int_{-1}^{1} f(x)dx$

$\int_{-0.2}^{1} f(x)dx$

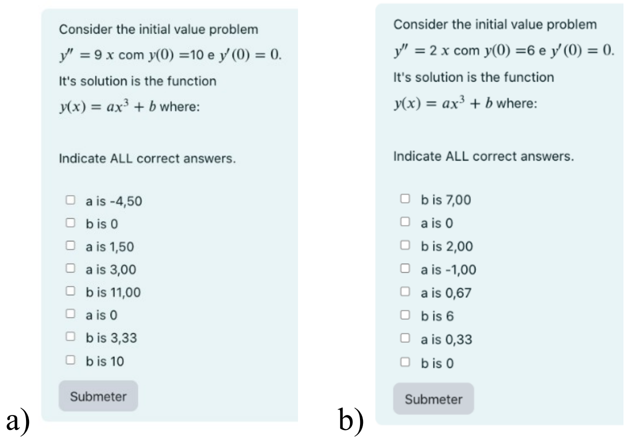
$\int_{-2}^{2} |f(x)|dx$

$\int_0^{1.5} \pi f(x)dx$

$\int_0^{1.5} -3f(x)dx$

Figure 1: a) The same quiz for the 4 classes, with the theme hidden for the students. b) and c) Quiz including a figure and mathematical symbols. Screenshot of a smartphone

Each quiz had just one question. Sometimes the questions are multiple choice like in Figure 2b) and c); other times they are numerical response as in Figure 2, and others have both as in Figure 3. Sometimes the question is the same for all students; other times it is an instance of the question with some randomly generated values, see Figure 2.



a)

Consider the initial value problem

$y'' = 9\,x$ com $y(0) = 10$ e $y'(0) = 0$.

It's solution is the function

$y(x) = ax^3 + b$ where:

Indicate ALL correct answers.

- □ a is -4,50
- □ b is 0
- □ a is 1,50
- □ a is 3,00
- □ b is 11,00
- □ a is 0
- □ b is 3,33
- □ b is 10

Submeter

b)

Consider the initial value problem

$y'' = 2\,x$ com $y(0) = 6$ e $y'(0) = 0$.

It's solution is the function

$y(x) = ax^3 + b$ where:

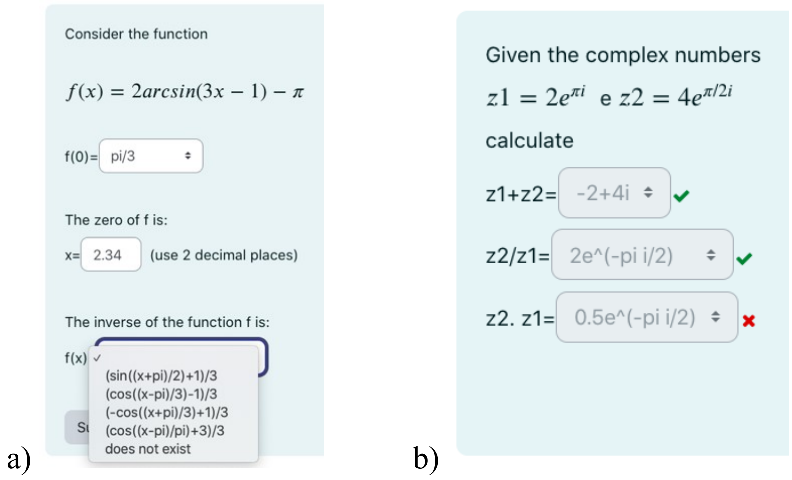
Indicate ALL correct answers.

- □ b is 7,00
- □ a is 0
- □ b is 2,00
- □ a is -1,00
- □ a is 0,67
- □ b is 6
- □ a is 0,33
- □ b is 0

Submeter

Figure 2: Two instances, aleatorily generated, of the same question, using numerical questions

By "submitting", students receive feedback, knowing which answers are correct and which are wrong. The next attempt already includes these answers; they just need to change them, see Figure 3.



a)

Consider the function

$f(x) = 2arcsin(3x - 1) - \pi$

f(0)= pi/3

The zero of f is:

x= 2.34 (use 2 decimal places)

The inverse of the function f is:

f(x) ✓

(sin((x+pi)/2)+1)/3
(cos((x-pi)/3)-1)/3
(-cos((x+pi)/3)+1)/3
(cos((x-pi)/pi)+3)/3
does not exist

b)

Given the complex numbers

$z1 = 2e^{\pi i}$ e $z2 = 4e^{\pi/2i}$

calculate

z1+z2= -2+4i ✓

z2/z1= 2e^(-pi i/2) ✓

z2. z1= 0.5e^(-pi i/2) ✗

Figure 3: A question with numerical and multiple choice questions. And the feedback.

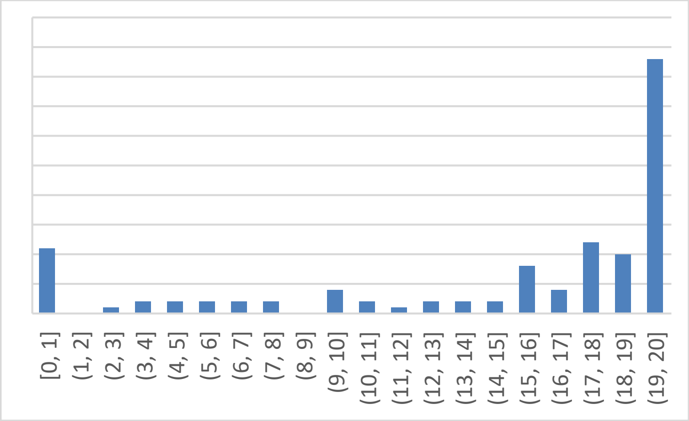
We were creative when making the questions allowing to evaluated all subjects, even those who seem difficult to evaluate using only numerical and multiple choice questions. The grades are all saved in Moodle, and it is very easy to export them.

Methodology, quantitative and qualitative data, and discussion

The research question is whether these quizzes are seen by students and teachers as an added value. To this end, a quantitative and qualitative study was carried out. The two research methods used were Survey and Interviews (Cohen, 2007). As tools, an anonymous survey, online in Moodle, was made available to all students on the course. And an Informal Conversational Interview of teachers regarding this new pedagogical practice was also collected.

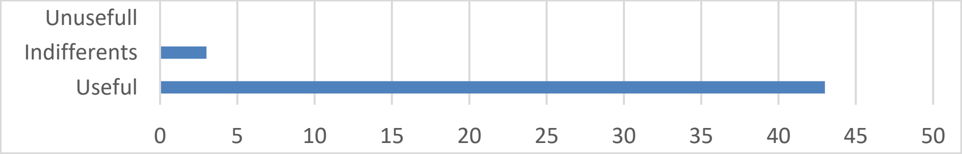
There was a total of 31 surprise quizzes in 39 classes of 1h30 each class. The course had 112 subscribed students, with 6 being considered ghost students since they never answered any class quiz, weekly quiz, test, or exam. Classes are not mandatory. The students who went to any class were 103. All 106 students (non-ghost) answered at least one quiz; of those, 96 students (91%) obtained a quiz average above 5 out of 20; 77 students (73%) got a quiz average above 15 and 43 students (41%)

obtained a quiz average above 19 out of 20 having answered almost all the quizzes with almost everything correct – as it is expected, since they may go to all classes/quizzes and when they have doubts solving, they can ask for help from the professor and from the classmates. This shows that the student participation rate was very high (see Graphic 1).



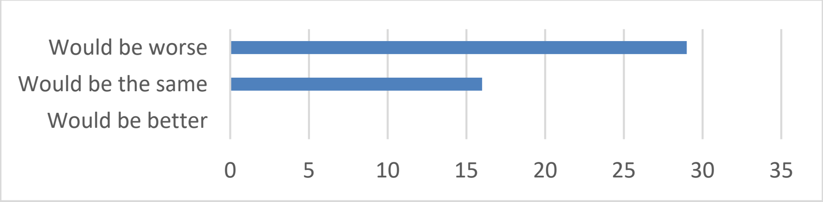
Graphic 1: Quizzes average grades histogram

The respondents to the survey were 47, representing 46% of the students who went to at least one class, with a representative distribution of student grades, including approved and failed students. About the question "Do you find the quizzes:" (Graphic 2), 43 find it useful, 3 indifferent and 1 didn't answer. So, 92% find it useful.



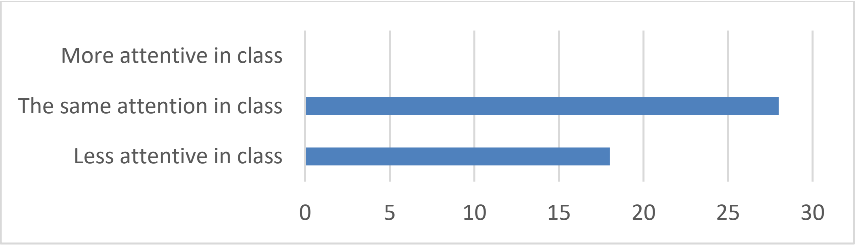
Graphic 2: Answers to the question "Do you find the quizzes:"

About the question "If there were no quizzes, do you think your grade would be different?" 29 think that it would be worse, 16 think that it would be the same and 2 didn't answer. Then, 62% believe that it makes them have a better grade (see Graphic 3).



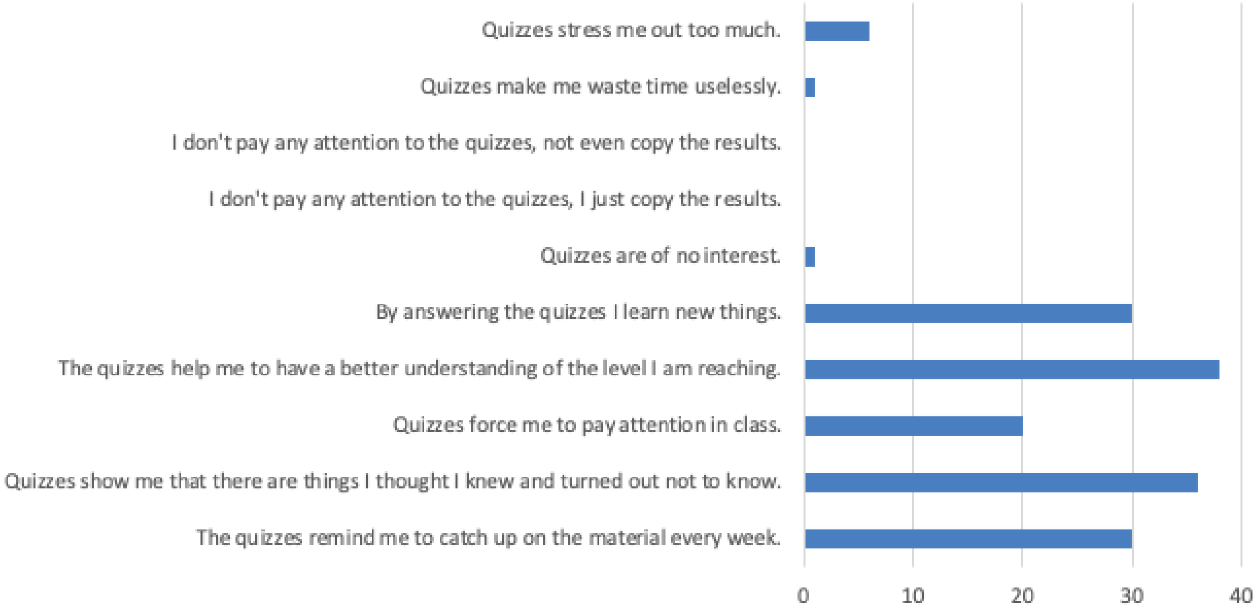
Graphic 3: Answers to the question "If there were no quizzes, do you think your grade would be different?"

About " If the quizzes didn't exist it would have been: more or less attentive in class" (Graphic 4), 18 think that without quizzes they would be less attentive in class, 28 think that would have the same attention and 1 didn't answer. Thus 38% believe that it makes them be more attentive (there are those who would already be attentive anyway, but, in principle, these are not counted).



Graphic 4: Answers to the question " If the quizzes didn't exist it would have been:"

Abou the next questions in Graphic 5, the number of respondents were (from top to bottom) 6, 1, 0, 0, 1, 30, 38, 20, 36, 30. In short, students believe that quizzes make them learn new things, study more, keep up with their studies, pay more attention, and be aware of the level they are reaching. Six students say that it gives them too much stress.



Graphic 5: Answers to the question " Please indicate ALL statements with which you agree:"

The open question of the survey was: “What do you like/dislike about quizzes?” we obtained 11 responses, all positive, reinforcing that it helps them to test whether they understand the material with basic exercises, also indicating that it forces them to pay attention in class and helps them learn more.

In an Informal Conversational Interview (Cohen, 2007) qualitative data was collected: the feedback of the two teachers of the course about this pedagogical practice. The teacher (who is the researcher) considers that the quizzes were a positive strategy. They made the students pay more attention in class, whenever she said “next comes a quiz”, the students studied more quietly and more attentively. It was interesting, that in one of the classes, when the quiz was launched, the students were all working hard, whether at the beginning of the year or until the end. In the other class, there were some who were disinterested. She believes that some students felt some stress, via that some were very concentrated and asked straight away if they didn't know how to solve a part and it was clear that they were anxious. Although there was no reason for that, because if the answer wasn't correct at the beginning, they were allowed to change it and there was never a lack of time: the teacher opened the test and only moved on to the next subject if almost everyone had already answered, and even let it open it until a bit after the end of the class. The teacher felt that it is time consuming, and sometimes is difficult to leave time for students to answer it.

The other teacher stated that quizzes are very important because they encourage all students to work and not leave studying until the end. He felt that they were important, because in his classes all the students worked when a quiz was launched, except for two students, who had a lot of difficulties, who pretended [note the pressure that quizzes generate] that they were solving it, and then ended up not handing it in. The teacher indicates that the time consumed by quizzes can be a problem, suggests to make shorter quizzes and/or fewer quizzes and that some might be done at home. Due to lack of time, he tried leaving some quizzes to finish at home but there were less students responding.

As discussion, first be aware that the surveys were automatically anonymous, meaning there was no pressure for students to respond positively to the survey, despite the researcher being their teacher. Student's participations rate in quizzes was very high with 72% of students obtaining 15 or more average values. And 41% over 19, that is, with almost all quizzes answered with full marks.

In the responses to the survey, an important result was that nearly all students find quizzes useful, 62% consider that the quizzes helped them to get a better grade, and 38% refer that were more attentive in class due to the quizzes. It should be noted that 6 students consider that the quizzes cause them a lot of stress. Students' opinion is that quizzes made them to learn new things, help them to have a better understanding of the level that they are reaching, make them pay more attention, and remind them to catch up on the subject.

Teachers also consider quizzes to be an effective way to make students more attentive in concordance with Nadeem & Al Falig (2020), to have a moment where they all work hard, and where they receive immediate feedback, again in concordance with literature (Booth,2012). In the future, given that quizzes take up a lot of class time, the intention is to move towards shorter questions or fewer quizzes.

Conclusions and future work

In order to create moments of active learning that at the same time provided immediate feedback to students, 31 online surprise quizzes were administered in class, in Moodle, around one per class, in a Calculus course to 112 students. Student participation in these quizzes was very high, answering almost all quizzes, with almost everything correct. In an anonymous online questionnaire with a very significant number of responses, student feedback was that almost everyone considered the quizzes to be useful. That makes them work harder, learn more and be more attentive. And that allows them to receive feedback and become more aware of the level they are reaching. It also makes them to be always updated.

Teachers also consider that quizzes are useful, make students more attentive and that quizzes also allow students to receive more constructive feedback as they can repeat until the result is correct. Due to the lack of class time, in the future we will have fewer quizzes or quicker quizzes to answer. Our findings are in line with literature: quizzes make students more attentive (Nadeem & Al Falig, 2020), and give immediate feedback which is positive for students (Booth,2012).

In short, given the high participation rate, the positive rating of the students and the encouraging reflection of the teachers, quizzes are undoubtedly a pedagogical strategy to maintain.

References

Beichner, R. J., Saul, J. M., Abbott, D. S., Morse, J., Deardorff, D., Allain, R. J., ... & Risley, J. S. (2007). The student-centered activities for large enrollment undergraduate programs (SCALE-UP) project. *Research-based reform of university physics*, *1*(1), 2–39.

Blanco, M., Estela, M. R., Ginovart, M., & Saà, J. (2009). Computer assisted assessment through moodle quizzes for calculus in an engineering undergraduate course. *Quaderni di Ricerca in Didattica*, *19*(2), 78–83.

Booth, D. J. (2012). Managing Mathematics with CALMAT. *International Journal of Innovation in Science and Mathematics Education (formerly CAL-laborate International)*, *2*(1).

Cohen, L., Manion, L. & Morrison, K. (2007). *Research methods in education* (6th ed.). Taylor and Francis Group. http://repository.unmas.ac.id/medias/journal/EBK-00127.pdf

Dori, Y. J., & Belcher, J. (2004). Improving Students' Understanding of Electromagnetism through Visualizations-A Large Scale Study. Paper submitted to the *2004 NARST Annual Meeting*, Vancouver, Canada. Online at http://web.mit.edu/jbelcher/www/TEALref/dori.pdf

Hill, M., Sharma, M. D., & Johnston, H. (2015). How online learning modules can improve the representational fluency and conceptual understanding of university physics students. *European Journal of Physics*, *36*(4), 045019.

Lasry, N., Mazur, E., & Watkins, J. (2008). Peer instruction: From Harvard to the two-year college. *American Journal of Physics*, *76*(11), 1066–1069.

Lim, L. L., Thiel, D. V., & Searles, D. J. (2012). Fine tuning the teaching methods used for second year university mathematics. *International Journal of Mathematical Education in Science and Technology*, *43*(1), 1–9.

Nadeem, N. H., & Al Falig, H. A. (2020). Kahoot! quizzes: A formative assessment tool to promote students' self-regulated learning skills. *Journal of Applied Linguistics and Language Research*, *7*(4), 1–20.

Shorter, N. A., & Young, C. Y. (2011). Comparing assessment methods as predictors of student learning in an undergraduate mathematics course. *International Journal of Mathematical Education in Science and Technology*, *42*(8), 1061–1067.

Siew, P. F. (2003). Flexible on-line assessment and feedback for teaching linear algebra. *International Journal of Mathematical Education in Science and Technology*, *34*(1), 43–51.

Twigg, C. A. (2005). Course Redesign Improves Learning and Reduces Cost. Policy Alert. *National Center for Public Policy and Higher Education*.

Wandel, A. P., Robinson, C., Abdulla, S., Dalby, T., Frederiks, A., & Galligan, L. (2015). Students' mathematical preparation: Differences in staff and student perceptions. *International Journal of Innovation in Science and Mathematics Education (CAL-laborate International)*, *23*(1), 82-93.

Assessment of students' learning mathematics with technology using video-based activities in an online course

Eirini Geraniou[1] and Cosette Crisan[2]

[1]IOE, UCL's Faculty of Education and Society, University College London, London, UK; e.geraniou@ucl.ac.uk

[2]IOE, UCL's Faculty of Education and Society, University College London, London, UK; c.crisan@ucl.ac.uk

Using technology for mathematical learning, but also for assessing students' mathematical learning has proven to enhance, support and impact mathematics education in innovative, yet challenging ways. One of the goals of the online asynchronous master's module we designed is to prepare postgraduate students (who are either prospective or practicing mathematics teachers) for assessing school students' mathematical learning when using digital technologies. Fostering postgraduate students' noticing and interpreting skills when analysing and assessing mathematical learning while a digital tool is used, has become a key priority for the design of our module's activities. This paper presents: (a) our current research study for investigating how best to support postgraduate students develop skills for assessing mathematical learning when using digital technologies; and (b) an innovative video-based activity that addresses this developmental need.

Keywords: Digital technologies, assessment, mathematical learning, video-based activity, professional development.

Introduction

Assessment plays a crucial role in the learning process, and in the digital era, we cannot underestimate the potential impact that digital technologies can have on mathematics education. Incorporating digital technologies into the assessment of mathematical learning opens up new possibilities, but also presents new challenges (e.g., Jankvist, et al., 2021; Drijvers & Sinclair, 2023). For instance, while automated assessments can offer immediate scoring, feedback, and adaptivity (e.g., Hoogland & Tout, 2018), there is a risk of overemphasising procedural fluency at the expense of capturing the depth and reasoning behind a student's response.

Over the past 15 years, there have been discussions about how 'slow' the transformation of assessment practices and policies in education with the support of digital technologies has been, despite the advancements in digital technologies (e.g., Timmis et al., 2016). For example, the rapid integration of Artificial Intelligence (AI) and tools like ChatGPT into educational assessments without proper research evidence and consideration of the implications is concerning. Therefore, it is important to gain an understanding based on rigorous research evidence, of how mathematical learning can be assessed when doing and learning mathematics with digital technologies. One of our goals as mathematics educators is to contribute to this research field whilst supporting the professional development of mathematics teachers in the digital era. In other words, we are carrying out a research study that investigates how best to support prospective and practicing mathematics teachers in assessing mathematical learning when a learner interacts with a digital tool. The context of our study is a ten-week online asynchronous master's module that introduces postgraduate students to several dynamic and interactive digital technologies for mathematical learning via numerous innovative activities. One such activity is the use of video-based activities, which are short videos that show school students working on mathematical tasks in GeoGebra and Desmos. Our module's postgraduate students are asked to analyse the videos and assess the learning of mathematics.

In this paper, we present some details about our research study, before moving on to describe the design and rationale of the innovative video-based activity. We conclude this short paper by discussing future outcomes and contributions of our study.

Research study and context

We are interested in identifying the best pedagogic strategies for supporting our postgraduate students, most of whom are prospective or practicing mathematics teachers, in developing skills for assessing school students' learning of mathematics while they interact with digital technologies. Before we go into more detail about the research work, we need to give a brief presentation of our master's module, which we refer to as the 'Digi' module.

The Digi module is taught online, with participants being given a series of tasks over a ten-week period. The weekly tasks are signposted on a virtual learning environment (Moodle) at the beginning of each week and include offline tasks such as: familiarisation with a piece of software and example problems using specific software, designing a maths activity using the specific digital environment, and trialling out the activity with learners. In our context, learners could be either school students from the schools that some of our postgraduate students (practising teachers) work at or school students from our postgraduate students' own personal networks. Online tasks include engaging with the ideas in the key readings of this module, reading one of the essential reading articles and writing a response about the points agreed or disagreed with from the article, and also contributing to online discussion forums with written observations on views and perspectives of their module peers. For example, in the third, sixth and ninth weeks, our postgraduate students are required to choose a software tool introduced in the prior two weeks, design a learning activity using features of good practice identified from the literature, use the activity they designed with students and analyse its implementation through engagement with research and the ideas assimilated from the literature reviewed to evaluate and justify the implications of using digital technology for students' learning. Being an asynchronous online course, our postgraduate students' contributions are solely in written format and consist of their weekly written task submissions, forum contributions such as written comments to peers' tasks, reflections on their own learning and peer assessed work, peer reviews and peer assessment.

The research study involves three student cohorts enrolled on our Digi module in 2024-2026 and it focuses on how students develop critically reflective and interpretative skills for assessing mathematical learning that takes place when a learner interacts with a digital tool. We will focus on different activities from the Digi module to answer the following main research question: *How are postgraduate students' skills for assessing mathematical learning that takes place when a learner interacts with a digital tool developed and supported by different activities?*, and our secondary research question: *In what ways do video-based activities develop postgraduate students' skills for assessing mathematical learning during interactions with a digital tool?*

Online video-based activity

To assist our postgraduate students in critically engaging with research and applying it to reflect on classroom practices involving digital technology, we are experimenting with the use of online video-based activities in one of the weekly activities. We have created videos that showcase pairs of school students actively engaging in mathematical tasks within a digital setting. We wanted to provide our postgraduate students with a simulation of a classroom-based scenario where two pupils worked together on a maths task involving digital technology. For this reason, the videos were not edited, and our students were invited to select their own segments of the recordings to analyse. Inspired by Van Es and Sherin's (2002) research, we explored the utilisation of video-based activities to offer our postgraduate students a shared learning episode for analysis. Video cases have been employed by numerous mathematics educators and researchers to guide teachers in focusing on students' learning and the decisions made by teachers during lessons. Van Es and Sherin (2002) suggested that videos could be effective tools in enhancing teachers' ability to observe, notice and interpret classroom interactions.

Among the various features of videos extensively documented in literature (Calandra et al., 2009; Van Es & Sherin, 2002), we highlight the capability of a video to be paused, rewound, and replayed multiple times, allowing viewers to focus on specific segments strategically chosen for their relevance to the viewers' goals, which in the case of the activity we present in this paper, is assessing school students' learning of mathematics during their interactions with Desmos and/or GeoGebra. The design of the video-based activity was guided by recommendations from researchers (Van Es & Sherin, 2002) emphasising that video clips could help shift attention away from teachers and classroom events, redirecting it toward students' work. In our research study, the videos produced are recordings of the collaborative efforts of a pair of school age students, narrowing the focus to the pedagogical activity of noticing significant episodes and analysing students' learning. The videos for this online module feature two Year 8 students, Tim and Tom (pseudonyms), both 12 years old, attending different secondary schools in a large city in the UK. Given the importance of understanding how students interacted with the provided digital environment, we utilized screencast video-recording software to capture on-screen work and audio recordings of student-student interactions during the mathematics activity. The maths activities presented to Tim and Tom were related to plotting points in a graphical environment that satisfy the equations of given straight lines; finding the equations of straight lines graphs already plotted; investigating and proving properties of quadrilaterals constructed in specific ways. Our postgraduate students were encouraged to watch these short videos and analyse how Tim and Tom used the digital tools to investigate the mathematics task. They were then invited to submit a piece of writing (800 words) where they assess and justify Tim and Tom's mathematics learning in a digital environment as portrayed by the videos.

It is worth adding that the ethical aspect of producing and utilising the videos underwent careful consideration (Flewitt, 2005). To ensure ethical standards, explicit consent from participants and parents was obtained, wherein the researcher transparently communicated the intended use of the video material and its purposes.

Concluding remarks

The first set of data collection from the first student cohort ends in March 2024 and we intend to present the findings of one of our enquiries: *In what ways do video-based activities develop postgraduate students' skills for assessing mathematical learning during interactions with a digital tool?* at the FAME1 conference. The data analysis will focus on postgraduate students' interpretations of the learners' interactions with GeoGebra and Desmos and assessment of potential learning outcomes, as observed in the video episodes, and reported in their written contributions. Evidence of development of critical reflection skills will also be sought in the postgraduate students' assignments where they are required to design, trial, evaluate and critically analyse a series of mathematical activities in their own practice, utilising the potential of digital technologies.

Future steps

In collaboration with our university partnership schools, the module tutors intend to produce a series of videos, with real classroom settings, featuring school students engaging in mathematics activities within a digital learning environment. These video recordings will then be edited to highlight key segments, transforming them into video-based activities for professional development. We are hoping to be able to offer guidance for the creation of such innovative video-based professional development activities with a focus on promoting and supporting effective assessment strategies when using digital technologies in mathematics learning.

References

Drijvers, P., & Sinclair, N. (2023). The role of digital technologies in mathematics education: purposes and perspectives. *ZDM - Mathematics Education.* https://doi.org/10.1007/s11858-023-01535-x

Flewitt, R. (2005). Conducting research with young children: some ethical considerations. *Early Child Development and Care, 175*(6), 553-565. https://doi.org/10.1080/03004430500131338

Hoogland, K., & Tout, D. (2018). Computer-based assessment of mathematics into the twenty-first century: Pressures and tensions. *ZDM - Mathematics Education, 50*, 675–686. https://doi.org/10.1007/s11858-018-0944-2

Jankvist, U. T., Dreyøe, J., Geraniou, E., Weigand, H.-G., & Misfeldt, M. (2021). CAS from an assessment point of view: Challenges and potentials. In A. Clark-Wilson, A. Donevska-Todorova, E. Faggiano, J. Trgalova, & H-G. Weigand (Eds.), *Mathematics Education in the digital age: Learning, Practice and Theory*, (pp.99–120). Routledge.

Timmis, S., Broadfoot, P., Sutherland, R., & Oldfield, A. (2016). Rethinking assessment in a digital age: opportunities, challenges and risks. *British Educational Research Journal, 42*(3), 454–476. https://doi.org/10.1002/berj.3215

Van Es, E., & Sherin, M. (2002). Learning to notice: Scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher Education, 10(*4), 571–596.

Examining ChatGPT responses to TPCK assessment items

Peter Gonschwerowski[1], Edith Lindenbauer[2] and Benjamin Rott[3]

[1]University of Cologne, Mathematics & Natural Sciences, Germany; pgonsche@uni-koeln.de

[2]University College of Education Upper Austria, Austria; edith.lindenbauer@ph-ooe.at

[3]University of Cologne, Mathematics & Natural Sciences, Germany; brott@uni-koeln.de

The development and, thus, the objective and valid assessment of the skill of selecting digital learning material (dLM) is essential for pre-service teachers. In this paper, we compare ChatGPT 3.5's responses with responses from pre-service teachers to items for assessing this skill to gain insights into ChatGPT's capabilities and the longevity of the items for its assessment. The results reveal that, for one, ChatGPT 3.5, so far, cannot analyze dynamic dLMs, and second, it does not make a decision on the use of the dLM but provides predominantly TPK but appropriate reasoning for using or not using it. ChatGPT's TPK responses are comparable to pre-service teachers' responses, and further studies are required to understand its impact thoroughly. Still, the presented results support the projected effects of ChatGPT on assessments in teacher education and the evaluation of the skill of selecting dLM.

Keywords: Artificial intelligence, performance-based assessment, teacher education, pre-service teachers, teacher evaluation.

Introduction

Selecting digital learning materials (dLMs) is a crucial skill for educators due to the possibilities dLMs offer in teaching and the varying quality of the many freely available dLMs. Therefore, this skill needs to be developed in teacher training, and valid and objective assessment instruments are required to assess the success of such development processes (König et al., 2022; Redecker & Punie, 2017). In this context, the most frequently used instruments so far are self-assessment instruments, which are based on the TPACK framework (Mishra & Koehler, 2006), and only a few valid and objective TPACK assessment instruments have been published, one of them using open-text items developed by Gonscherowski et al. (in review). With the free public availability of text-based natural language processing and multimodal artificial intelligence (AI) models, examining such assessment instruments and respective items becomes increasingly important because we must understand how AI chatbots respond to such items and potentially undermine their assessment purpose. Therefore, this paper examines the ability of ChatGPT 3.5 to respond to items developed and validated in Gonscherowski et al. (in review) for assessing the (pre-service) teachers' skill of selecting dLM by having to reason for or against the use of a given dLM for specific learning content in the context of a specific learner age and potential special educational needs.

Existing multimodal AI systems process text and images and either generate new images, modify images, or categorize images (Livberber & Ayvaz, 2023). However, such systems do not currently interpret dynamic dLMs and their functionality or intended learning goal. We chose ChatGPT 3.5, although it is a text-based AI model and cannot directly analyze or interpret images because of its popularity and free use (Livberber & Ayvaz, 2023). We developed a textual description of a dLM and its functionality and provided it to ChatGPT to answer the following research questions.

RQ1: How does ChatGPT reason for or against using a dLM for a specific group of learners, characterized by their age and special learning needs, when provided with a text-based description of said dLM?

RQ2: How does ChatGPT's reasoning compare to that of pre-service teachers who evaluated the same dLM?

By answering these research questions, we want to understand the longevity of the assessment items and gain insights on how to refine them in the future so that ChatGPT or similar tools cannot undermine an assessment of pre-service teachers using the items.

Theoretical framework: TPACK

To assess the skill of selecting dLM, we rely on the TPACK framework by Mishra & Koehler (2006). The framework is frequently used to describe and assess the digital competence of pre-/in-service teachers (Gonscherowski & Rott, 2023). It describes the content knowledge (CK), pedagogical knowledge (PK), and technological knowledge (TK) that teachers require to integrate technology into teaching successfully. In the framework, pedagogical content knowledge (PCK), technological content knowledge (TCK), technological pedagogical (TPK), and technological pedagogical content knowledge (TPCK) detail the knowledge required because of the interplay of CK, PK, and TK. Mishra and Kohler (2006) define TK, TCK, TPK, and TPCK as follows:

> TK is knowledge about standard technologies, such as books, chalk and blackboard, and more advanced technologies, such as the Internet and digital video. (ibid., pp. 1027–1028)
>
> TCK is knowledge about the manner in which technology and content are reciprocally related. Although technology constrains the kinds of representations possible, newer technologies often afford newer and more varied representations and greater flexibility in navigating across these representations. (ibid., p. 1028)
>
> TPK is knowledge of the existence, components, and capabilities of various technologies as they are used in teaching and learning settings, and conversely, knowing how teaching might change as the result of using particular technologies. (ibid., p. 1028)
>
> TPCK [...] is the basis of good teaching with technology and requires an understanding of the representation of concepts using technologies; pedagogical techniques that use technologies in constructive ways to teach content. (ibid., pp. 1028-1029)

We refer to Mishra and Kohler (2006, pp. 1026-1027) for the definitions of CK, PK, and PCK. The descriptions of TCK, TPK, and TPCK entail reasons for using dLM in a teaching setting and can also be used to categorize reasons for or against using a dLM in a teaching situation.

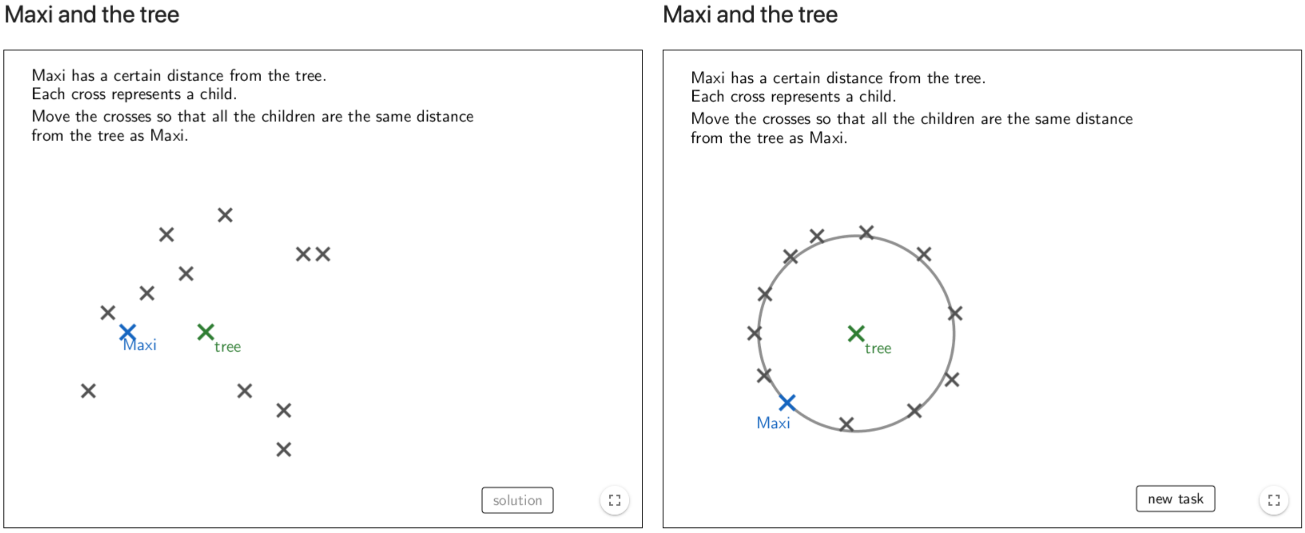
Method

We apply a qualitative case study method to answer the research questions. To do so, we compare ChatGPT 3.5's responses to items one to four (see Table 4) developed by Gonscherowski et al. (in review), with the responses by pre-service teachers when assessing the skill of selecting dLM using a specific dLM (see Figure 1). First, we outline the coding and scoring of the items, and then we compare ChatGPT and pre-service teachers' results.

Assessing the skill of selecting dLM

In Gonscherowski et al. (in review), four items were developed and validated to assess the skill of selecting dLMs (see Table 4). These items aim to evaluate pre-service teachers' understanding of the learning content that a particular dLM is intended to deliver (one open-ended item), the learner age group (grades 1-13, in two-year increments), and the special learner needs with which the dLM would be used (two closed items). A fourth open text item inquires about the reasons for or against using the dLM.

To evaluate the items' ability to assess the skill of selecting dLM reliably, validly, and objectively, they were integrated into an online test using a specific dLM (see Figure 1). The online test was distributed among mathematics pre-service teachers, and participation was voluntary.



Maxi and the tree

Maxi has a certain distance from the tree.
Each cross represents a child.
Move the crosses so that all the children are the same distance from the tree as Maxi.

Maxi and the tree

Maxi has a certain distance from the tree.
Each cross represents a child.
Move the crosses so that all the children are the same distance from the tree as Maxi.

Figure 1: Starting and end point of the dLM

There were 379 participants each from one university in Germany (*n = 314*) and one in Austria (*n = 65*). The German pre-service teachers were distributed in programs on primary education (*n = 110*), special education (*n = 173*), and lower secondary education (*n = 31*); the Austrian pre-service teachers were enrolled in a combined program for lower and upper secondary levels. Furthermore, the participants covered all relevant semesters of the mathematics education program: first year (*n = 57*), second year (*n = 149*), third year (*n = 93*), and year seven or higher (*n = 80*).

The responses to the items were coded following qualitative content analysis and summarized in scores: items one to three were analyzed with scores from zero to three, and item four on a scale from zero to four. Table 1 shows example responses for items one to three for a generic and detailed description of the learning content appropriate for the learner age group and the special learner needs; the former scored two, and the latter scored three. Inadequate descriptions of the learning content and generic or detailed descriptions of the learning content but that were inappropriate for the selected learner age group or the selected special needs were scored zero and one, respectively.

Table 1: Example responses by pre-service teachers (items one to three)

Definition of the score (items 1 to 3)	Example response
A generic description of the learning content appropriate for learner age/special educational needs	"Introduction to the definition circle' Learner age: 3-4 and hearing and communication learning needs" (WS2022/2023_ANJA07B_Pre)
A detailed description of the learning content and appropriate for learner age/special educational needs	"To derive the definition of a circle line. All points on the circumcircle have equal distance from the center.' Learner age: 5-6 no special educational needs" (WS2022/2023_EDJO06B_Pre)

The responses to item four, "Justify why or why not you would use the digital material," were coded and scored using the codebook developed based on the TPACK framework (Mishra & Koehler, 2006) and other validated and generally accepted studies. The reasons for or against the use of dLM were categorized as TCK, TPK, and TPCK-based arguments. Coding of the TCK reasoning entails arguments such as different or dynamic representations of the learning content and reducing or increasing extraneous cognitive load because of the dLM. The coding of the TPK reasoning encompasses arguments such as self-directed learning, differentiation, learner motivation, exploring, teacher efficiency, and learner distraction.

Table 2: Single TCK and TPK example reasoning by pre-service teachers (item four)

Example response 1	Example response 2
TCK: "My students (with special needs) would be overwhelmed with the dLM..." (WS2022/2023_MAHA10L_Pre)	TPK: "Because by moving the crosses around, the children can discover for themselves what properties all the points of a circle have." (WS2022/2023_DOMA15S_Pre)

The example responses shown in Table 2 were scored as two, containing a single TCK or TPK reason for or against using the dLM. Responses with either two TCK or TPK reasons were scored three, and responses with both TCK and TPK, thus TPCK reasoning, were scored four, see Table 3.

Table 3: TCK and TPK, thus TPCK example reasoning by pre-service teachers (item four)

Example response 1	Example response 2
"...provides a different way of practicing and illustrating the properties of a circle. However, it is not necessarily suitable for all learners with special needs. The tree and the learning atmosphere outside the classroom (not in the usual environment) can lead to too much distraction." (WS2022/2023_CHGU24M_Pre)	"...Changing environments (digital/non-digital) with students and exploring other learning environments encourages interest. Experiencing on their own how these mathematical relationships are connected makes understanding easier." (WS2022/2023_SAHE17B_Pre)

Example one in Table 3 entails "...a different way of practicing and illustrating...", a TCK argument, and "...lead to too much distraction...", a TPK argument, thus constituting TPCK reasoning. No reasoning was scored zero, and generic arguments were scored one.

The coding and the scoring of the suitability of the dLM for a particular learning age, learning needs, and learning content were derived from the local curriculum (items one to three). The arguments for or against the dLM (item four) were coded based on the codebook as outlined. For the skill of selecting dLM, the scores were combined on a scale from zero to seven.

The development process of a textual description of the dLM for ChatGPT

The dLM presented in Figure 1 encompasses the mathematical topic of a circle. Learners should discover the concept of the circle based on its defining property: a figure consisting of all the points in the plane that are a specific distance (radius) from a certain point (the center). Since ChatGPT applies a language-processing AI model and only accepts text input, the three authors crafted a textual description of the dLM using collaborative editing and multiple review cycles. In the processes of crafting the description, the authors applied the following guiding rules: a) explicitly state that the description is one of dynamic learning material, b) describe the activity the learners need to perform, c) not use the term circle or properties of the circle, and d) use as much of the wording of the task description included in the dLM as possible. The process resulted in the following description of the dLM and its dynamic functionality, including an introduction. "The dynamic learning material we want to evaluate presents a task in a dynamic digital applet showing a tree, a child named Maxi, and further children represented by crosses. Maxi is at a pre-set distance from the tree. The learners should move the crosses representing the children so they all have the same distance from the tree as Maxi. Finally, learners can press a button named "solution," and the solution to the task is revealed." ChatGPT was provided with the description and the four assessment items.

In Table 4, abbreviated ChatGPT responses are contrasted with example responses of pre-service teachers. The full transcript of the ChatGPT session and the text description of the dLM were recorded (OpenAI, 12/25/2023). ChatGPT responses are coded and scored as outlined in the previous section (see Tables 1, 2, and 3).

Results

ChatGPT's answer to item one, the learning content the dLM is intended to provide, is incorrect as the dLM is not intended for spatial reasoning in geometry or possibly proportional reasoning (see Table 4). The correct learning content contains the characteristics of a circle as a shape consisting of all points of equal distance from a given point, the center. The assigned grade level by ChatGPT (early to middle stage of primary education) is also incorrect. Although the circle is covered in the middle stages of primary school, the concept of equal distance from the center to the circle's circumference is only covered in lower secondary school, according to the curriculum in Germany and Austria. ChatGPT's responses to item three, the special needs of learners, contain multiple impairments. Each impairment is described with a sentence, of which "fine motor skills," "cognitive challenges," and "attention challenges" are appropriate. However, ChatGPT suggested that the dLM could be used with auditory learners if the dLM would be auditive narrated, which was not indicated in the description of the dLM. Although correct, it is conjecture and inappropriate.

The scoring as outlined previously resulted in a score of zero pt. for items one to three. Within the text responses, ChatGPT references elements of the description of the dLM, e.g., “Maxi and other children represented by the crosses” or “solution button suggests [...] immediate feedback,” providing specificity and context to its responses, however not all are suitable in response to the items.

The justifications for using or not using the dLM, required in item four provided by ChatGPT, are appropriate but predominantly correspond to TPK reasoning, namely “engaging and interactive” (motivation), “immediate feedback” (self-directed learning), the potential for “inclusivity” or “sensory overload” (both differentiation), and “motor skill challenges” (inclusion). Other reasons provided are coded as overly general arguments (e.g., “overreliance on technology” and “lack of flexibility”) and incorrect argumentation, such as “concept reinforcement,” as the dLM is for the discovery of the learning content and not for its reinforcement. ChatGPT references elements of the text description of the dLM, for example, “arranging crosses around the tree...” and “...button providing immediate feedback...” to support its arguments.

Table 4: Responses to items assessing the skill “selecting dLM” by pre-service teachers and ChatGPT

#	Items	Example responses of pre-service teachers	pt	ChatGPT 3.5 responses	pt
1	Describe the learning content for which you think the digital material is intended.	“To derive the definition of a circle line. All points on the circumcircle have equal distance from the center.” (WS2022/2023_EDJO06B_Pre)	3	“...it seems that the dynamic learning material is designed for teaching and practicing concepts related to spatial reasoning, geometry, and possibly proportional reasoning in a mathematics education context....”	0
2	For which learner age do you think the digital material is suitable?	“5-6 grade” (WS2022/2023_EDJO06B_Pre)		“...the digital material seems to be designed for learners in the early to middle stages of primary education...”	
3	In your opinion, is the digital material suitable for learners with special educational needs? If so, which ones?	“No special educational needs.” (WS2022/2023_EDJO06B_Pre)		“...the dynamic learning material seems to have characteristics that could make it inclusive and adaptable for a range of learners, including those with certain special educational needs: visual learners, auditory learners, fine motor skills, cognitive challenges, customization and adaptability, attention challenges...”	

4	Justify why or why not you would use the digital material.	"...provides a different way of practicing and illustrating the properties of a circle. However, it is not necessarily suitable for all learners with special needs. The tree and the learning atmosphere outside the classroom (not in the usual environment) can lead to too much distraction." (WS2022/2023_CHGU24M_Pre)	4	"Certainly, the decision to use the digital material depends on the specific educational context, learning objectives, and the needs of the learners. Here are justifications both for and against using the digital material: Reasons to use the dLM: Engaging and interactive, visual learning, immediate feedback, concept reinforcement, potential for inclusivity; Reasons not to use the dLM: motor skill challenges, overreliance on technology, sensory overload, lack of flexibility."	3

Without making an actual decision on the use of the dLM, ChatGPT's reasoning for and against using the dLM corresponds to a score of three points, as previously outlined.

Regarding RQ1, "How does ChatGPT reason for or against using a dLM," our analysis reveals that ChatGPT does not adequately capture the learning content and the learner age range of the learning content. This is not surprising, as the specifics of the curriculum for one differ by local and are not necessarily consistent. Advances in ChatGPT responses in this regard are expected, particularly with the ChatGPT 4.0 personalized memory feature, which allows localized information like the curriculum to be added to a ChatGPT user profile. Further, ChatGPT 4.0 does not require a textual description of the dLM. In addition, ChatGPT does provide generic and predominant correct TPK reasoning to justify using or not using the dLM.

Regarding RQ2, "How does the reasoning of ChatGPT compare to pre-service teachers who evaluated the same dLM?" Table 4 reveals that, for item four, ChatGPT's responses are comparable to those shown in Tables 2 and 3. However, ChatGPT does not decide and only provides reasons for and against using the dLM. The combined score for the entire sample of pre-service teachers (n = 379), as presented in Table 5, shows a total mean score of 2.18 (SD = 1.59), which underscores the good results of ChatGPT (score of 3) and echoes its reported capabilities and impact on assessments in higher (teacher) education (Livberber & Ayvaz, 2023).

Table 5: Scores of pre-service teachers and ChatGPT

mean; SD; max score	pre-service teachers (*n* = 379)	ChatGPT
Items one-three (learning content)	0.93; 1.06; 3.00	0.00
Item four (reasoning)	1.25; 1.03; 4.00	3.00
Σ representing the score for "selecting dLM"	2.18; 1.59; 6.00	3.00

Particular to note is ChatGPT's capability of providing subject-unspecific (TPK) reasoning and its lack of TCK reasoning, which needs to be leveraged when using the four items and a dLM for assessing the skill "selecting dLM." One should use a dLM that is complex enough to provide various arguments, and we recommend dLMs for discovering (mathematical) concepts, as such a dLM enables potentially more argumentation, specifically more TCK arguments, as, for example, a dLM for drill and practice.

Outlook and Limitations

As ChatGPT 3.5 cannot examine dynamic dLM, an inherent limitation of the case study lies in the authors' description of the dLM. Higher-quality responses by ChatGPT could have potentially been achieved with a more elaborate description of the dLM and by further refinement of the items to cater to ChatGPT. However, we did not want to optimize the responses of ChatGPT, but rather the opposite, as one can conjecture if pre-service teachers can precisely describe the functionality of a dLM and inquire ChatGPT in a way to achieve high-quality responses, they may also possess the skill of selecting dLMs we want to assess with the items. The use of dLM for evaluating the skill of selecting dLM increases the technology requirements of the assessment. In environments with limited online access, the risk posed by using ChatGPT may also be lower. In the future, further testing with other and newer AI models (ChatGPT 5.0) and dLMs is required to understand the impact of assessing the skill with the developed items thoroughly. In addition, the responses of ChatGPT should be compared with responses by in-service teachers whose responses potentially exceed those of pre-service teachers, as hypothesized in Gonscherowski and Rott (2022).

References

Gonscherowski, P., Lindenbauer, E., Kaspar, K., & Rott, B. (in review). Selecting digital learning material as an approach to assess pre-service teachers' digital competence.

Gonscherowski, P., & Rott, B. (2023). Selecting digital technology: A review of TPACK instruments. In M. Ayalon, B. Koichu, R. Leikin, L. Rubel, & M. Tabach (Eds.), *Proceedings of the 46th conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 378–386). PME.

Gonscherowski, P., & Rott, B. (2022). How Do Pre-/In-Service Mathematics Teachers Reason for or against the Use of Digital Technology in Teaching? *Mathematics, 10*(13), 2345. https://doi.org/10.3390/math10132345

König, J., Heine, S., Jäger-Biela, D., & Rothland, M. (2022). ICT integration in teachers' lesson plans: A scoping review of empirical studies. *European Journal of Teacher Education*, 1–29. https://doi.org/10.1080/02619768.2022.2138323

Livberber, T., & Ayvaz, S. (2023). The impact of Artificial Intelligence in academia. *Heliyon, 9*(9). https://doi.org/10.1016/j.heliyon.2023.e19688

Mishra, P. & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

OpenAI. (12/25/2023, 4:29:14 PM). ChatGPT [Large language model] https://chat.openai.com. https://chat.openai.com/share/fce6ba96-b89a-4802-9093-04fee07bee5b

Redecker, C., & Punie, Y. (2017). European framework for the digital competence of educators – DigCompEdu. *Publications Office of the European Union*. https://doi.org/10.2760/159770

Open-ended items in digital formative assessments: Decision Trees as (AI-compatible) approach to reliably code students' understanding?

Corinna Hankeln

TU Dortmund University, Dortmund, Germany; corinna.hankeln@math.tu-dortmund.de

Open-ended items, in which students draw images, explain meanings or argue, allow them to express their own mental representations of situations and make it possible to grasp even fragile concepts in nuances and details. However, those answers-types are rarely found in digital formative assessment, also because they are often difficult to evaluate. This paper reports on the integration of open-ended items into the digital formative assessments of the Mastering Math – Online-Check and exemplifies for an item on conceptual understanding of multiplication how current approaches of category-based scoring could be optimized by using decision trees to rate features of responses. In preparation for the integration of an automatic pre-coding by an artificial intelligence, an exploratory study is presented on the functioning of prompt-based classification of students' answers by ChatGPT.

Keywords: Evaluation methods, Digital formative assessment, conceptual understanding, decision tree, AI-prompts.

Open-ended items are "worth the effort" in digital formative assessments

Formative assessments have the potential to promote the implementation of conceptual learning in classrooms (Burkhardt & Schoenfeld, 2018). However, many digital formative assessment (DFA) platforms hold a dominant procedural focus (Hoogland & Tout, 2018), partly because procedural items are easier to (automatically) evaluate. In order to uncover shallow understanding and to assess deep conceptual understanding, items are needed where students translate a concept between different (e.g., verbal, graphical, symbolical, or contextual) representations, explain the meaning of particular concept elements and the connection between representations, or connect different concept elements in a wider network of elements (Hiebert & Carpenter, 1992). There are proofs of existence that students' thinking can be assessed by well-designed multiple-choice formats (e.g., the SMART test, Stacey et al., 2018), but open-ended long-answer or complex graphical formats give students more opportunities to express their own mental representations of situations (without being influenced by distractors), allowing thus the demonstration even fragile concepts in details and nuances (Hankeln et al., submitted). Furthermore, students' language production for describing mathematical structures or explaining meanings are relevant learning goals (Götze & Baiker, 2021; Prediger, 2022) that should not be excluded in assessments, also because those responses provide valuable resources for subsequent communication processes between students and teachers.

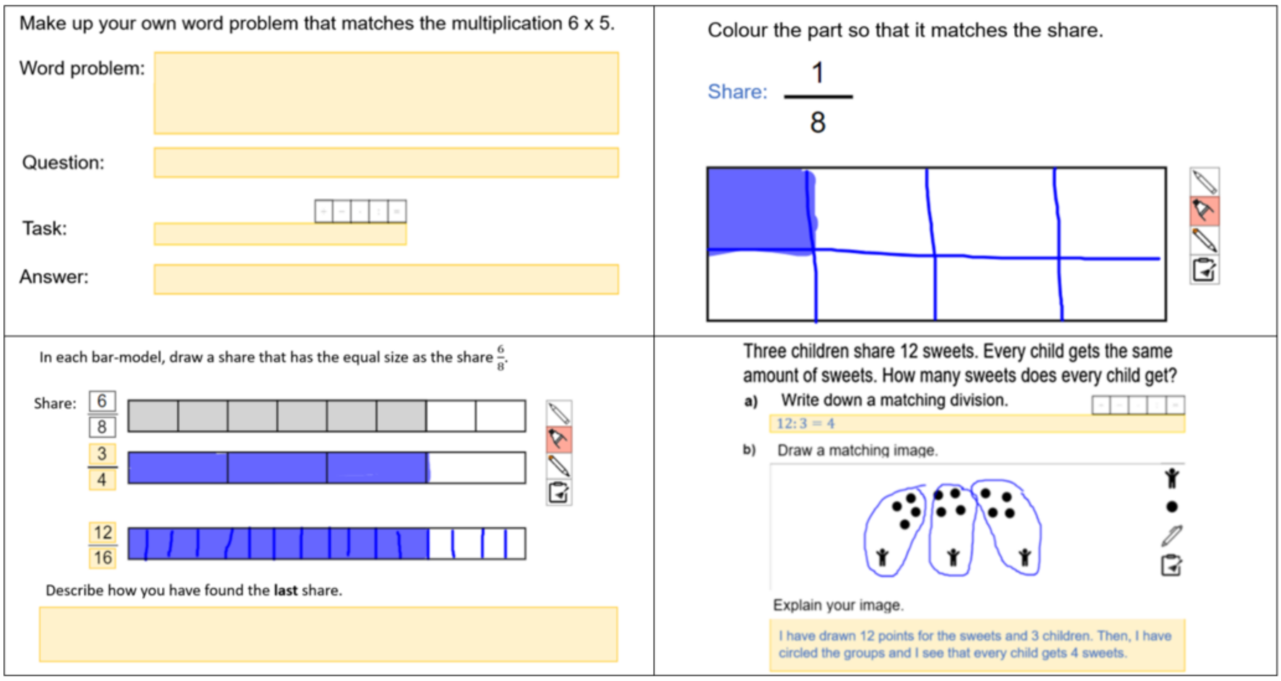
Typical challenges in coding open-ended items

Well-designed open-ended items come with the price that the evaluation of those responses requires topic-specific epistemic background knowledge, taking into account the current position of students' learning progression, knowledge about the relevant components of the assessed topic, like concept elements, representations, and language needed to explain them (Siemon, 2019) and typical misconceptions. In their meta-study of 14 DFA tools, Çekiç and Bakla (2021) state

> "As for open-ended items, no fully reliable methods of grading have been created so far, but there have been significant developments in this area. Several tools have put an effort in developing systems to grade open-ended items. There have been four methods of grading: (1) autoscoring of short-response questions […], (2) auto-grading based on the existence of a set of pre-determined keywords […] (3) assigning numerical scores manually […] and (4) the use of artificial intelligence for scoring open-ended items. Each of these methods is valuable in a time when we desperately need ways to deal with open-ended responses. […] Obviously, the success of the keyword method or artificial intelligence is open to debate and should be tested empirically, yet they seem to be good starting points for further developments." (p.1477)

This paper presents a small-scale exploratory study to address this research gap, contrasting a combination of (2) and (3), namely manually classifying open-ended items based on different coding schemes with (4), the use of few-shots prompts to ChatGPT to code students' responses, all within the DFA Mastering Math – Online-Check.

Open-ended items in the Mastering Math – Online-Check



Make up your own word problem that matches the multiplication 6 x 5.

Word problem:

Question:

Task:

Answer:

Colour the part so that it matches the share.

Share: $\frac{1}{8}$

In each bar-model, draw a share that has the equal size as the share $\frac{6}{8}$.

Describe how you have found the **last** share.

Three children share 12 sweets. Every child gets the same amount of sweets. How many sweets does every child get?

a) Write down a matching division.

12 : 3 = 4

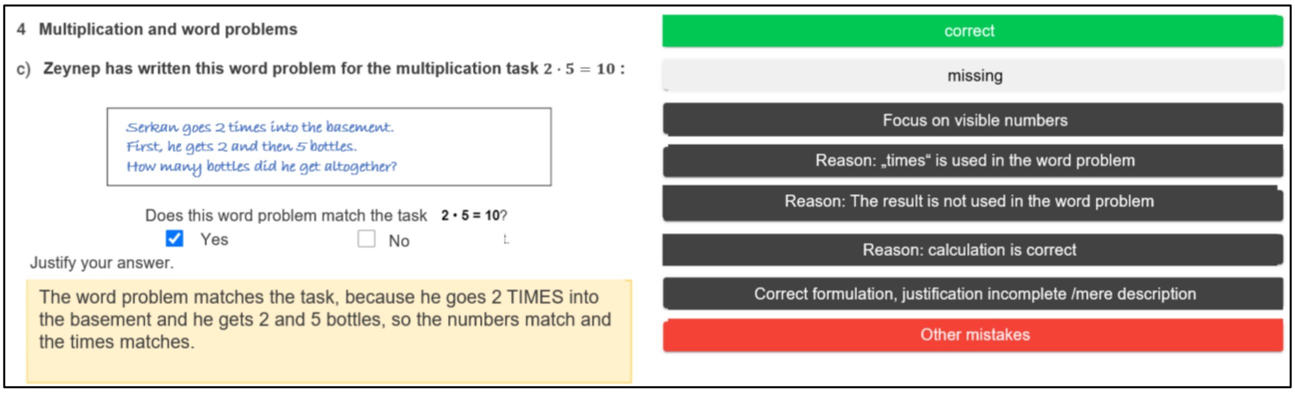
b) Draw a matching image.

Explain your image.

I have drawn 12 points for the sweets and 3 children. Then, I have circled the groups and I see that every child gets 4 sweets.

Figure 1: Exemplary open-ended items in the Mastering Math Online-Check

The Mastering Math – Online-Check (Hankeln et al., submitted) is currently developed as a DFA that is integrated in the 15-year-long Mastering Math project which aims at Grade 5–7 (10- to 13-year-old) students who struggle in mathematics and need a second learning opportunity for understanding basic arithmetic concepts such as the place value understanding or meanings of multiplication and division (Prediger et al., 2019). Each of the 45 Online-Checks is linked to teaching material, and the results of every Online-Check provides support for the prioritization of learning tasks and communicative prompts in remediation classes. The Online-Checks are administered in the newly created platform *alea.schule*. When teachers have chosen an Online-Check for their students in this platform, students can access the assessment via any browser on a tablet or computer. When students have filled out an Online-Check, their answers get send to the teacher-platform *alea.schule*. All items in closed formats (multiple-choice or single-choice items, short answers, drag-and-drop answers, etc.) are automatically coded as correct or incorrect regarding typical misconceptions. Open answers need to be manually coded by teachers, supported by suggested item-specific categories entailing typical solutions and errors (Figure 2). So the items can be evaluated not just if they are right or wrong, but "wrong in a specific way" (Stacey et al., 2018, p. 246). The evaluation outcomes can be displayed in different evaluation dashboards with varying degrees of details and focus. The Online-Check thus aims at informing teachers to support their planning of subsequent lessons and does not provide any direct feedback to students.



4 Multiplication and word problems

c) Zeynep has written this word problem for the multiplication task $2 \cdot 5 = 10$:

Serkan goes 2 times into the basement.
First, he gets 2 and then 5 bottles.
How many bottles did he get altogether?

Does this word problem match the task $2 \cdot 5 = 10$?

☑ Yes ☐ No

Justify your answer.

The word problem matches the task, because he goes 2 TIMES into the basement and he gets 2 and 5 bottles, so the numbers match and the times matches.

correct

missing

Focus on visible numbers

Reason: „times" is used in the word problem

Reason: The result is not used in the word problem

Reason: calculation is correct

Correct formulation, justification incomplete /mere description

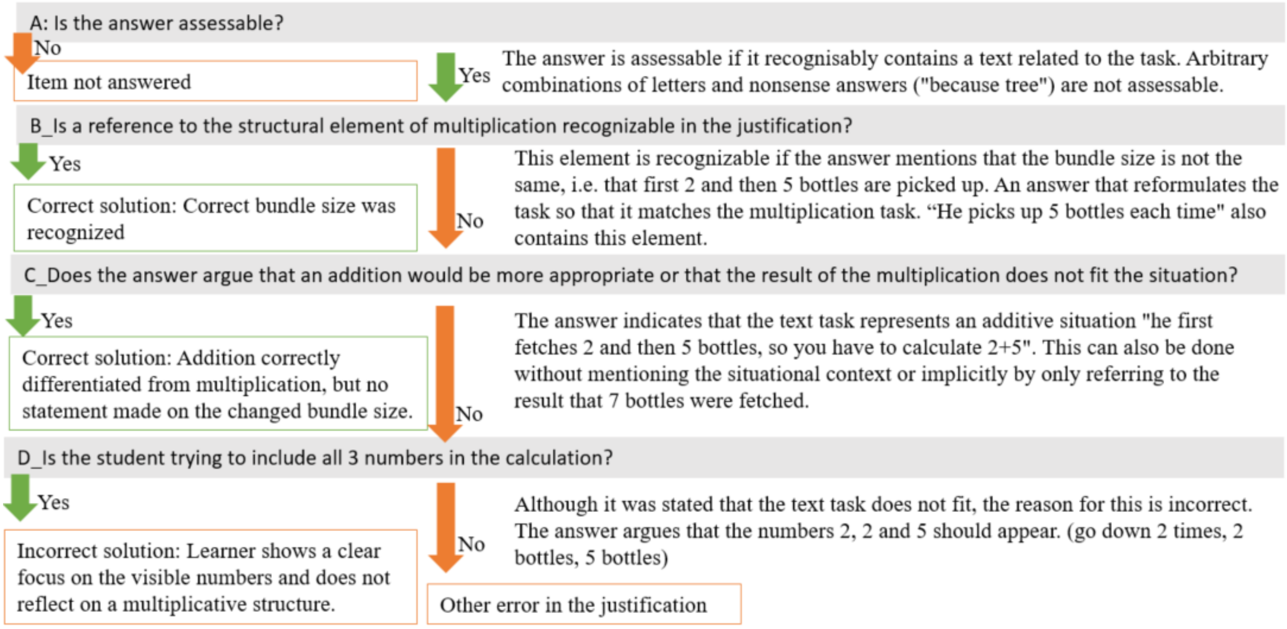
Other mistakes

Figure 2: Coding area in the platform *alea.schule*: Category-based coding to evaluate of students' responses

Evaluation of category-based scoring and proposition of decision trees

A necessary (but insufficient) precondition for the validity of the conclusions drawn from the classification of students' responses is that the coding of open items is reliable (Çekiç & Bakla, 2021). However, a pilot study with 15 pre-service teachers (in their mathematics teacher master program) who coded 50 responses to the item "Bottles" in Figure 2 following the proposed categories revealed low interrater-reliabilities (3 rater per student response, $\kappa = .21$).

There can be various reasons for this observation: The correct choice of a category is highly dependent on the raters' pedagogical content knowledge (Prediger et al., 2023). Without an accurate understanding of the categories, raters cannot identify central indicators for these categories within student responses. Whereas research projects overcome this challenge by detailed rater preparation, teachers – in their daily use of the tool – need to be able to code different items without detailed instruction. That is why proposed buttons for selecting categories have to be labeled precisely, taking into account frequent misconceptions. While there are ideas how to improve the comprehensibility, for example by including category descriptions, another approach is to integrate a feature-based scheme to evaluate the answers in form of decision trees (Kingsford & Salzberg, 2008). Students' responses are thus seen as texts that have to labelled, which makes the coding of students' answers to a form of text-classification problem (Gasparetto et al., 2022). There are various approaches to text-classification as it is widely used for example in spam-filters or website-classification, and one of them is decision trees. A decision tree is a sequence of questions about features associated with the items (Kingsford & Salzberg, 2008). The questions thereby form a hierarchy, encoded as a tree. There are statistical means to design such hierarchical trees, for example to ensure that the data is divided into groups with similar variances by each questions (Kingsford & Salzberg, 2008). However, for evaluating students' responses, the hierarchy is derived from the goal of the assessment and has to be grounded in topic-specific mathematics education backgrounds.



A: Is the answer assessable?

Item not answered

The answer is assessable if it recognisably contains a text related to the task. Arbitrary combinations of letters and nonsense answers ("because tree") are not assessable.

B_Is a reference to the structural element of multiplication recognizable in the justification?

Correct solution: Correct bundle size was recognized

This element is recognizable if the answer mentions that the bundle size is not the same, i.e. that first 2 and then 5 bottles are picked up. An answer that reformulates the task so that it matches the multiplication task. "He picks up 5 bottles each time" also contains this element.

C_Does the answer argue that an addition would be more appropriate or that the result of the multiplication does not fit the situation?

Correct solution: Addition correctly differentiated from multiplication, but no statement made on the changed bundle size.

The answer indicates that the text task represents an additive situation "he first fetches 2 and then 5 bottles, so you have to calculate 2+5". This can also be done without mentioning the situational context or implicitly by only referring to the result that 7 bottles were fetched.

D_Is the student trying to include all 3 numbers in the calculation?

Incorrect solution: Learner shows a clear focus on the visible numbers and does not reflect on a multiplicative structure.

Although it was stated that the text task does not fit, the reason for this is incorrect. The answer argues that the numbers 2, 2 and 5 should appear. (go down 2 times, 2 bottles, 5 bottles)

Other error in the justification

Figure 3: Decision tree for the Item "Bottles" (from Figure 2)

Moons (2023) reports that some teachers consider grading schemes to be more efficient even when in fact it does not accelerate their grading process. A grading scheme is a set of statements from which teachers can select those that match the students' response (also called check-box grading). As the traditional, holistic grading in this study already had a very high interrater-reliability ($\kappa > 0.8$), a significant improvement by introducing the check-box grading could only be observed for one item.

For the Item "Bottles" (Figure 2), we designed a decision tree in order to evaluate students' conceptual understanding of multiplication (Figure 3). In this paper, we analyze those answers that justify their (correct) decision that the word problem posed by a fictitious student does not match the multiplication 2 x 5. The first step in the decision tree was to identify "nonsense" answers that cannot be used for an assessment,. The second question aimed at identifying if answers indicate students' understanding that the posed word problem uses counted units of different size which does not correspond to multiplication unit structures (first 2 bottles, then 5 bottles instead of two times the same amount of bottles). This is the essential aspect of the item to capture if a student uses expressions referring to the bundle sizes. If not, the next question checks if the answer argues that the word problem would fit an addition (and (implicitly) argues that it thus cannot be a multiplication). All while being correct as well, this answer reveals insight into students' conceptual understanding of the contrast between addition and multiplication. Differentiating between these nuances allows to make detailed diagnoses of individual students to enchain specific follow-up questions only to certain students. If this third question is answered negatively, the response-text is checked if the student correctly rejected the proposed multiplication but based his decision on incorrect reasons. As the posed word problem contains three numbers, a typical surface strategy is to blindly take and combine them. So, the third question tries to identify those answers that draw upon this surface strategy by saying for example “the second 2 is missing in the calculation, it has to be 2 x 2 x 5 = 20”. All remaining response-texts are assumed to be atypical mistakes that cannot unambiguously be related to theoretical misconceptions.

Empirical Study: Human and AI-coding with a decision tree

Methods

To investigate if the decision tree-based approach is a suitable way to optimise the coding of open-ended items, 124 children’ responses to the Item “Bottles” from a pilot study of the Online-Check were coded with the decision tree. Firstly, two trained raters coded all responses independently, discussed differences and decided on a final coding (“expert-rating”) that is used as base-line to compare the quality of other ratings. Secondly, 15 pre-service teachers in their master studies coded a subset of 50 responses according to the decision tree. Every pre-service teacher received 10 responses to code in a rotated design. Those 10 responses were compiled to be a representative set of answers in order to avoid systematic misunderstandings of the questions biasing the ratings. The sample was drawn from a stochastics course for second year pre-service teachers at TU Dortmund University, which did not relate to the topic of the item. There were no additional information provided for the raters other than those in Figure 3. This coding resulted in a dataset (“teacher-rating”) with three ratings per students’ response. The different codings were compared (a) within the group of pre-service teachers in order to estimate their agreement (using Fleiss Kappa) and (b) between the expert-rating and the teacher-rating. Thirdly, the AI ChatGPT was asked with the help of few-shot prompts to classify the students’ responses analogically. This coding was also compared to the expert-rating and the teacher-rating based on the accuracy (proportion of predictions that are correct), the precision (proportion of positive predictions that are correct) and the sensitivity (recall) (proportion of positive answers that are correctly predicted).

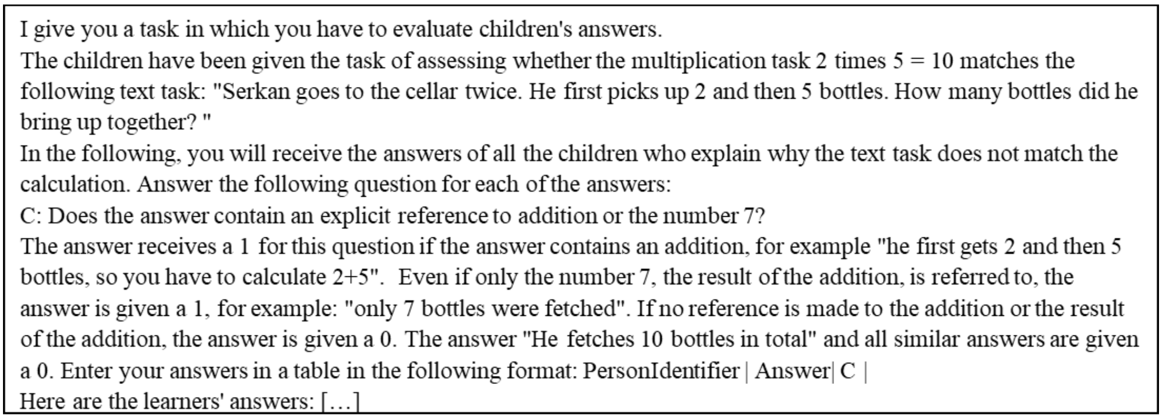
Findings

The expert-rating revealed that even though the hierarchical structure is only developed with respect to the assessed content, every group of responses is represented (Table 1) and only 38 responses (31 % of all responses) belonged to the “other error” category. 21 responses (55 % of this category) showed atypical mistakes like misunderstanding the situation (“he gets two and five bottles and he does that two times, so it has to be two times seven”), the others gave incomplete justifications (“he goes two times in the basement”) or gave no reasons at all (“it does not fit”).

	End of decision-tree branch	Continuation of decision-tree
A	No: 5 responses (5 %)	Yes: 119 responses (96 %)
B	Yes: 33 responses (27 % of remaining responses)	No: 86 responses (73 % of remaining responses)
C	Yes: 39 responses (45 % of remaining responses)	No: 47 responses (55 % of remaining responses)
D	Yes: 9 responses (19 % of remaining responses)	No: 38 responses (81 % of remaining responses)

Table 1: Distribution of responses categories

The “teacher-rating” conducted by master students showed an improvement in the interrater-reliability for the coding based on a decision tree compared to the classical category-based approach (see above), ranging between κ = .41 (B), κ = .63 (C) and κ = .58 (D). It is interesting to see that the question that requires the most pedagogical content knowledge about unitizing or multiplication as counting in groups is the category with the lowest agreement. This question had an insufficient agreement between expert-rating and teacher-rating. The pre-service teachers only coded 59 % of the responses like the expert-rating for question B, while 73 % of agreement was reached for question C. Question D showed the lowest agreement with 39 %. It has to be kept in mind that the pre-service teachers did not receive any examples or explanations for the different questions.



I give you a task in which you have to evaluate children's answers.
The children have been given the task of assessing whether the multiplication task 2 times 5 = 10 matches the following text task: "Serkan goes to the cellar twice. He first picks up 2 and then 5 bottles. How many bottles did he bring up together? "
In the following, you will receive the answers of all the children who explain why the text task does not match the calculation. Answer the following question for each of the answers:
C: Does the answer contain an explicit reference to addition or the number 7?
The answer receives a 1 for this question if the answer contains an addition, for example "he first gets 2 and then 5 bottles, so you have to calculate 2+5". Even if only the number 7, the result of the addition, is referred to, the answer is given a 1, for example: "only 7 bottles were fetched". If no reference is made to the addition or the result of the addition, the answer is given a 0. The answer "He fetches 10 bottles in total" and all similar answers are given a 0. Enter your answers in a table in the following format: PersonIdentifier | Answer| C |
Here are the learners' answers: […]

Figure 4: Prompt to Chat-GPT to answer question C in the decision-tree (Figure 3)

In order to explore how a Large-Language-Model such as ChatGPT can evaluate the responses without being a priori trained with labelled data, we formulated few-shot prompts, where we described the item (classifying students' responses), the origin of the data (students' responses to the item "Bottles") and explained the questions that ChatGPT had to answer for every response (Figure 4). To improve the quality of the coding, we also included a few examples for the decisions yes or no respectively. Those examples were given both as general description and with a precise example. We iterated the prompt design and revised the prompts when we could identify systematic misunderstandings. We report here the statistics of the best fitting prompts. The identification of non-rateable responses worked very well with an accuracy of 96 %. In two cases, ChatGPT found a rateable response to be non-rateable, so the precision was a 100 % but the recall (sensitivity, how well a yes-answer can be detected) was 98 %. This error would thus lead to the abort of the coding process and the loss of diagnostic information. The identification of the structural element of the multiplication was identified in 32 cases in the expert-rating. All of those cases have also been identified by the AI, the recall was thus 100%. However, 21 cases have been falsely diagnosed to make reference to the structural element of multiplication (precision: 60 %). In total the accuracy was 82 %. The expert-rating revealed 56 cases where no reference to the structural element was made but a reference to the addition or the result of the addition. 50 of these cases have also been detected by ChatGPT (recall 89 %), 11 cases were falsely marked (precision 82 %). The accuracy was 80 %. For Question D, that asks if students decided correctly but due to an incorrect surface-strategy, the expert-rating identified nine cases of the remaining 31 responses. Six of them were found by the AI (recall 67 %), five answers were falsely accused of a surface strategy (precision 55 %) and the accuracy was at 65 %. All questions showed that with the few-shots prompt, the recall (sensitivity) was higher than the precision, meaning that the identification of true yes-answers works well with the price that there are several false positive classifications. For Questions B and C, this would imply that the problem is falsely classified as correct and possible problems are not detected. For Question D this implies that the surface strategy is more often suspected than true. The balance of both error types is of course a challenge, but for the specific use of the formative assessment, we would prefer, especially for Question B, to have a higher precision.

Decision tree question	n	accuracy $\left(\frac{correct\ predictions}{all\ predicitions}\right)$	recall $\left(\frac{true\ positive\ predictions}{true\ positives + false\ negatives}\right)$	precision $\left(\frac{true\ positive\ predictions}{positive\ predicitions}\right)$
A rateable answer?	124	95.7%	98.3%	100%
B structural element of multiplication?	118	82.2%	100 %	60.4 %
C addition?	87	80.5 %	89.3 %	82.0 %
D surface strategy?	31	64.5 %	66. 7 %	54.6 %

Table 2: Accuracy, recall and precision of ChatGPT's classification of decision tree questions (Figure 3)

Discussion and conclusion

Open-ended items are challenging to use in any assessment, but they bring enormous advantages especially for formative assessments aiming at capturing students' conceptual understanding in details and nuances (Hankeln et al., submitted). This small, exploratory study gave insight into the challenges of the evaluation of open-ended items and proposed the use of decision trees to get a precise impression of the features of a response while on the same time improving the reliability of a scoring. The empirical findings show that interrater-reliability of pre-service teachers can indeed be improved by a question-based decision tree. However, in this non-representative sample, it did not reach a satisfactory level. This can of course be due to insufficient topic-specific epistemic background knowledge of the pre-service teachers who have not yet finished their studies, but this could also indicate the need for additional information on the expected coding, also when a decision tree is used. Such additional information could either be general descriptions that can be accessed on demand, or exemplary codings, like they were included in the few-shot prompt that was given to Chat-GPT. Our first results seem to confirm Çekiç and Bakla (2021), that AI-based coding is a promising approach for future development. In our case, however, we saw a tendency of ChatGPT to have a better recall than precision. This has to be investigated further, especially in contrast to other AI-based classifier like for example BERT.

Acknowledgment

Funding. The Mastering Math Online-Check is developed and investigated within the projects SchuMaS (funded by the National Ministry of Education and Research, grant No. SchumaS-SMS2101L-01PR2101C to S. Prediger) and Mastering Math (funded by the Departments of Education, Hamburg and Hessia to S. Prediger, C. Hankeln, and U.Kroehne).

References

Çekiç, A., & Bakla, A. (2021). Review of digital formative assessment tools: Features and future directions. *International Online Journal of Education and Teaching*, *8*(3), 1459–1485.

Gasparetto A, Marcuzzo M, Zangari A, Albarelli A. (2022). A survey on text classification algorithms: from text to predictions. *Information*. *13*(2), 83. https://doi.org/10.3390/info13020083

Götze, D. & Baiker, A. (2021). Language-responsive support for multiplicative thinking as unitizing: Results of an intervention study in the second grade. *ZDM – Mathematics Education*, *53*(2), 263–275. https://doi.org/10.1007/s11858-020-01206-1

Hankeln, C., Kroehne, U., Voss, L., Gross, S. & Prediger, S. (submitted). Developing digital formative assessment for deep conceptual learning goals: Which topic-specific research gaps need to be closed? Submitted manuscript.

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D. A. Grouws (Eds.), *Handbook of research on mathematics teaching and learning* (pp. 65–97). Macmillan.

Hoogland, K. & Tout, D. (2018). Computer-based assessment of mathematics into the twenty-first century: Pressures and tensions. *ZDM – Mathematics Education*, *50*(4), 675–686. https://doi.org/10.1007/s11858-018-0944-2

Burkhardt, H. & Schoenfeld, A. (2018). Assessment in the service of learning: Challenges and opportunities or Plus ça Change, Plus c'est la même Chose. *ZDM – Mathematics Education*, *50*(4), 571–585. https://doi.org/10.1007/s11858-018-0937-1

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature Biotechnology*, *26*(9), 1011–1013. https://doi.org/10.1038/nbt0908-1011

Moons, F. (2023). Semi-automated assessment of handwritten mathematics tasks: Atomic, reusable feedback for assessing student tests by teachers and exams by a group of assessors. [Doctoral thesis 3 University of Antwerp]. https://hdl.handle.net/10067/1980770151162165141

Prediger, S. (2022). Enhancing language for developing conceptual understanding. In J. Hodgen, E. Geraniou, G. Bolondi, & F. Ferretti (Eds.), *Proceedings of 12th Twelfth Congress of the European Society for Research in Mathematics Education* (pp. 8–33). University of Bolzano / ERME.

Prediger, S., Dröse, J., Stahnke, R., & Ademmer, C. (2023). Teacher expertise for fostering at-risk students' understanding of basic concepts: Conceptual model and evidence for growth. *Journal of Mathematics Teacher Education, 26*(4), 481–508. https://doi.org/10.1007/s10857-022-09538-3

Prediger, S., Fischer, C., Selter, C., & Schöber, C. (2019). Combining material- and community-based implementation strategies for scaling up: The case of supporting low-achieving middle school students. *Educational Studies in Mathematics*, *102*(3), 361–378. https://doi.org/10.1007/s10649-018-9835-2

Siemon, D. (2019). Knowing and building on what students know: The case of multiplicative thinking. In D. Siemon, T. Barkatsas, & R. Seah (Eds.), *Researching and Using Progressions (Trajectories) in Mathematics Education* (pp. 6–31). Brill.

Stacey, K. Steinle, V., Price, B. & Gvozdenko, E. (2018). Specific mmathematics assessments that reveal thinking: An online tool to build teachers' diagnostic competence and support teaching. In T. Leuders, K. Philipp, & J. Leuders (Eds.), *Diagnostic Competence of Mathematics Teachers* (pp. 241–261). Springer.

Validity objections to comparative judgement

Ian Jones

Loughborough University, Department of Mathematics Education, UK; I.Jones@lboro.ac.uk

Comparative judgement approaches to assigning grades to students' work have received interest from mathematics education researchers over recent decades. These approaches involve assessors deciding which of two presented pieces of work is 'better', and the decisions are then converted into scores. Several objections have been raised to using comparative judgement for summative assessment and in this theoretical paper I respond to objections that such approaches are 'not valid'. These include objections that there is a lack of evidence supporting validity, and that researchers assert comparative judgement is 'intrinsically valid' in ways that are incomplete and inconsistent. I argue that most validity objections are addressed by published evidence, and that the validity of applying comparative judgement to mathematics education assessments is a special case.

Keywords: Comparative judgement; alternative assessment; summative assessment; validity.

Comparative judgement

Comparative judgement approaches to generating scores or grades from students' exam scripts or test responses have been gaining attention in education over recent decades (Bartholomew & M. Jones, 2022; I. Jones & Davies, 2023). A key motivation for using comparative judgement approaches to assess mathematical learning is to encourage a shift away from the types of short, objective items that are rife in examinations and tests (Swan & Burkhardt, 2012). Short, objective items threaten consequential validity if we value mathematics students performing longer-form mathematical activities such as problem solving, sustained reasoning and explaining their understanding (NCTM Research Committee, 2013). Importantly, comparative judgement approaches can encourage the shift away from short, objective items because they readily reduce assessor inconsistency compared to marking for the case of open-ended and relatively unstructured assessment tasks across a range of mathematical topics and student age ranges (e.g. Davies et al., 2020; Hunter & I. Jones, 2018; I. Jones & Inglis, 2015).

The approach applies Thurstone's (1927) well-established Law of Comparative judgement which states that human beings are poor at making absolute judgements (e.g. determining the temperature in a room in Celsius) and good at making relative judgements (e.g. determining if it is colder in the room or outside). Thurstone's law can be applied to educational assessment by harnessing the collective judgement of a group of subject experts (Pollitt, 2012). In practice, an online platform presents pairs of students' written mathematical work ('responses') to expert assessors via an internet browser. Interested readers can try judging pairs of responses to the test prompt "What is an equation?" here: http://tinyurl.com/190523equation. Once many judgements have been collected from several assessors the binary decision data is converted into a unique score for each response. Due to space constraints I do not detail the mathematics or other statistical details here, but see Pollitt (2012) and I. Jones and Davies (2023) for technical details.

Comparative judgement approaches differ from marking because there are no rubrics, no scoring of short items, and no aggregation of scores into a single mark or grade. Instead, outcomes are grounded in several assessors' holistic judgement of direct evidence of student work. Validity has been conceptualised for comparative judgement assessments in terms of the "collective understanding of the construct by a relevant community of experts" (I. Jones & Inglis, 2015, p. 341). Validity has been empirically evaluated across mathematics education studies using a range of constructs including criterion, divergent and content validity (e.g. Bisson et al., 2019). Study foci have included undergraduate modules on calculus (I. Jones & Alcock, 2014), statistics (Bisson et al., 2016) and proof comprehension (Davies, Alcock & I. Jones. 2020), and secondary school topics including fractions (I. Jones et al., 2013), statistics (Marshall et al., 2020) and problem solving (I. Jones & Inglis, 2015). Studies have also investigated conceptual understanding across a range of topics at secondary level (I. Jones & Karadeniz, 2016) and primary level (Hunter & I. Jones, 2018). Across these studies and others the outcomes of using comparative judgement have been found to be robust in terms of assessor consistency (called 'reliability' here on; Verhavert et al., 2019), and in terms of validity (Bartholomew & M. Jones, 2022; I. Jones & Davies, 2023).

The validity objections

Most of the objections that have I encountered have not been published in the scholarly literature but have arisen in other sources such as social media, blog posts, conversations at conferences or seminars, and reviewers' comments on submitted articles. I focus here on what I perceive to have been the most common, and I strive to present them coherently and fairly.

A common objection is that there is a lack of evidence in support of the validity of comparative judgement for summative assessment. For example, van Daal et al. (2022) wrote, in the context of education generally, "only a limited number of studies dig into the validity of comparative judgement" (p.2). Others go further, suggesting there is something inherently challenging about validity when it comes to comparative judgment (e.g. Bokhove, 2019). A particular concern is the opacity of assessment decisions lacking detailed criteria, which it has been alleged means that comparative judgement "drains from the assessment any considerations of *what* [*sic.*] is being assessed & the whole question of validity worth having" (Davis, 2017).

Perhaps the most substantive objections to the validity of comparative judgement approaches, including for mathematics education, came from a theoretical paper by Kelly, Richardson and Isaacs (2022). Before I present and respond to the key objections presented in their paper, it is worth clarifying that the vast majority of literature on using comparative judgement for educational assessment is authored by advocates, of which I am one. This has resulted in a body of literature overtly committed to demonstrating its virtues with few dissenting voices. Kelly et al.'s paper is therefore a needed and refreshing counter to the growing pro-comparative judgement corpus, to which the current paper adds, and it would be healthy for the discipline if further sceptical researchers publish scholarly critiques.

Kelly et al. critiqued what they called the "intrinsic validity" (p.2022) rationale for using comparative judgement. The intrinsic validity rationale is that validity can be defined in terms of what experts collectively deem 'good' answers to be and therefore, advocates of comparative judgement argue, outcomes are inherently valid. Above, I quoted this as the "collective understanding of the construct by a relevant community of experts" (I. Jones & Inglis, 2015, p. 341). Kelly et al. offer four specific critiques of the intrinsic validity rationale: (i) there is no evidence relative judgements are superior to absolute judgements; (ii) researchers rarely define 'expert'; (iii) non-experts, specifically students comparatively judging peers' work, produce outcomes that correlate well with experts' outcomes; (iv) researchers sometimes remove poorly performing ('misfitting') experts thereby undermining their very definition of validity. They argue that "it is inconsistent to justify the use of a method based on its psychological underpinnings, and also to contend that the details of this theory are irrelevant provided the method works in practice" (p.679).

Finally, an objection related to validity is that comparative judgment can only produced norm-referenced grades (e.g. Dolan, 2021). Norm-referenced grading involves allocating grades statistically, such as awarding the top 10% of scores a grade A. This means that grade A for a 'weaker' cohort is not equivalent to grade A for a 'stronger' cohort. In contrast, criterion-referenced grading involves allocating grades against agreed standards. This means that we would expect fewer grade A's to be awarded to the 'weaker' cohort than the 'stronger' cohort (Lok, McNaught & Young, 2016). It has been asserted that comparative judgement can only be used for norm-referenced grading, presumably because each student's performance is judged relative to the other students' performances.

Responding to the validity objections.

It is the case that there could and should be research into the validity of comparative judgement for summative assessment, including across different mathematical topics, student age ranges, jurisdictions and so forth. However, most of the assertions of a lack of evidence seem not to acknowledge, let alone critique, the evidence that has been published over the past decade. In fact, and as mentioned above, numerous studies have evaluated constructs such as the convergent, divergent and face validity of comparative judgement assessments.

Convergent validity has been demonstrated by correlating or predicting outcomes with independent measures (standardised instruments, specially designed tests, achievement data, teacher estimates) across age ranges, learning contexts and mathematical topics (e.g. I. Jones et al., 2016; I. Jones et al. 2019; Marshall et al., 2020). For example, Bisson et al. (2019) conducted a randomised controlled trial using two outcome measures of students' conceptual understanding of calculus – one based on comparative judgement and the other based on a traditional standardised test – and compared the results.

Divergent validity has been investigated using independent measures we would not expect to correlate with or predict comparative judgement-based mathematics outcomes. For example, Bisson et al. (2019) found that secondary students' mathematics but not English GCSE grades predicted comparative judgement scores (GCSE refers to a terminal school qualification in parts of the United Kingdom). In another example, I. Jones et al. (2013) used comparative judgement to assess secondary students' explanations of how to put fractions in order. They found that conceptual but not procedural measures of mathematical knowledge predicted comparative judgement scores, thereby demonstrating the potential of comparative judgement to reliably assess conceptual understanding rather than procedural knowledge. Other studies have shown that non-expert judges (peers, lay people) produce comparative judgement outcomes that diverge from those of experts (e.g. I. Jones & Alcock, 2014; I. Jones & Wheadon, 2015).

Content validity has been demonstrated using various techniques including expert review (e.g. I. Jones & Inglis, 2015), thematic analysis and related coding methods (e.g. Davies et al. 2021), and interviewing or surveying judges (e.g. Hunter & I. Jones, 2018; I. Jones et al., 2014; Marshall et al., 2020). An increasingly common method is to qualitatively code students' responses using existing frameworks (e.g. I. Jones & Karadeniz, 2016) or grounded approaches (e.g. Davies et al., 2020), and then use multiple regression techniques to identify the features of high-scoring responses. For example, I. Jones and Karadeniz (2016) applied a published coding scheme (Hunsader et al., 2014) and found that quantity written, use of numbers, and use of graphics predicted comparative judgement scores, but other features such as use of letters or use of 'real-world' examples did not. Similarly, Davies et al. (2020) developed a grounded coded scheme for undergraduates' definitions of proof, and found that comparative judgement scores produced by research mathematician judges were consistent with typical characterisations of proof reported by philosophers of mathematics (Davies et al., 2020).

This body of evidence partly addresses Kelly et al.'s objection that researchers assume that comparative judgement is intrinsically validity. We have seen convergent, divergent and content validity has been investigated across a variety of contexts, and not merely assumed to be inherent, the case for summative assessment in mathematics education.

I now turn to Kelly et al.'s specific objections of the intrinsic validity rationale.

(i) Kelly et al. claimed that researchers have not established that relative judgements are superior to absolute judgements for producing reliable outcomes. However this is not the case, at least for the case of secondary school peer assessment. I. Jones and Wheadon (2015) showed that for absolute judgement outcomes, inter-rater reliabilities were effectively zero (mean $r = -.02$), whereas the relative (comparative judgement) outcomes, reliabilities were high (mean $r = .86$).

(ii) It is fair to claim that researchers tend not to provide a precise or universal definition of 'expert'. It is also fair to counter that expertise is clearly defined within the context of many published studies, such as "mathematics PhD students" (Davies et al., 2020, p. 188). Experts are also sometimes contrasted against non-experts such as peers or novices (e.g. I. Jones & Alcock, 2014). Nevertheless, these operationalisations and contrasts tend to be buried in methods sections, and researchers could and should be more upfront and explicit about the term 'expert'.

(iii) A related point is Kelly et al.'s objection that non-experts' outcomes sometimes "correlated well" (p. 682) with the outcomes of experts. This has indeed been the case in some studies (e.g. I. Jones & Alcock, 2014; I. Jones & Wheadon, 2015), and moreover has been used as an argument for peer judgements contributing to summative outcomes. However, there is an important caveat: studies typically use many times more peer judgements than expert judgements because the former are easier to obtained. Crucially, the number of judgements per student answer collected is correlated with reliability (see Verhavert et al., 2019, for a detailed explanation and demonstration), and this largely explains peers' robust outcomes. Moreover, when novices (e.g. participants who have received no mathematics education beyond secondary school and have not studied calculus) make comparative judgements of undergraduate mathematicians' responses to an open-ended calculus question, correlations are lower still (e.g. I. Jones & Alcock, 2014).

(iv) Kelly et al.'s objection that removing 'misfitting' experts undermines claims that validity is intrinsic to collective expert judgement has merit. Opinions as to whether or not misfitting experts should be identified and removed varies across comparative judgement researchers in my experience. I am of the view that removing misfitting experts is generally not necessary or desirable (see I. Jones & Davies, 2023) bar rare occasions when an expert appears not to have judged in good faith, or with adequate attention (e.g. they completed their judgements suspiciously quickly). Therefore, Kelly et al. make a good theoretical point and researchers should consider carefully the rationale and validity of removing misfitting expert judges.

Finally, it is not the case as asserted by some (e.g. Dolan, 2021) that comparative judgment can only be used to produce norm-referenced grades. There are several methods are available for criterion-referencing comparative judgment scores when producing grade boundaries. For example, Marshall et al. (2020) included exemplar grade-boundary scripts in the judging pot of students' responses. The boundary scripts' scores were then used as cut-scores to assign grades. In fact comparing the standards of grades across different cohorts is a particular strength of comparative judgement approaches. For example, I. Jones et al. (2016) used comparative judgement to investigate changes over five decades of a terminal qualification (A-level Mathematics) in England. Conversely, albeit in the context of creative writing, Heldsinger and Humphry (2013) used comparative judgement to identify grade boundary scripts, which were then disseminated to teachers in order to establish consistent writing standards in Australia.

Discussion.

I have presented objections to the validity of using comparative judgement to produce summative scores or grades. I have argued that in the main these objections are addressed by the published research evidence. That is not to say more validity evidence would be unwelcome. For example, recent developments have included eye-tracking equipment to investigate how experts make judgement decisions of argumentative writing (Gijsen et al., 2021), and such methods should be applied in the context of mathematics education. In addition, while criterion-referencing methods have been applied to comparative judgement outcomes (e.g. Marshall et al., 2020), these methods have not been widely published or, to the best of my knowledge, systematically investigated.

I agree with Kelly et al.'s conclusion that there is a need for a "comprehensive, systematic review of the evidence to explore to what extent the rationales for using comparative judgment have empirical support" (p. 684). Critical review is essential to scholarly progress and, based on the theoretical and detailed objections of Kelly et al., comparative judgement researchers should consider giving greater thought to and being clearer about their use of the term 'expert'. They should also reflect carefully on the implications for validity, including Kelly et al.'s critique of what they call intrinsic validity, when considering removing misfitting expert judges.

Final comment

Objections to the validity of comparative judgement tend not to distinguish between different subjects. There is a motivation to use comparative judgement that is specific to *mathematics* education which I set out in the introduction to this paper: comparative judgement can reduce exam and test designers' dependence on short, objective test questions, and allow greater use of open-ended, relatively unstructured formats, without decreased assessor consistency (i.e. reliability). This is not the case for other subject areas. For example, researchers have investigated using comparative judgement to assess students' creative writing (e.g. Wheadon et al., 2020). Here the motivation, unlike for assessing mathematics, is not to enable different types of test questions, but to improve the reliability of assessing existing test questions (Pollitt, 2012).

To conclude, Kelly et al.'s "intrinsic validity" could be amended for the case of mathematics education to include the theoretical argument that comparative judgement enables the inclusion of relatively open and unstructured items in exams and tests, and therefore can increase their validity with no loss of reliability.

References

Bartholomew, S. R., & Jones, M. D. (2022). A systematized review of research with adaptive comparative judgment (ACJ) in higher education. *International Journal of Technology and Design Education, 32*, 1159–1190. https://doi.org/10.1007/s10798-020-09642-6

Bisson, M.-J., Gilmore, C., Inglis, M., & Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education, 2*(2), 141–164. https://doi.org/10.1007/s40753-016-0024-3

Bisson, M.-J., Gilmore, C., Inglis, M., & Jones, I. (2019). Teaching using contextualised and decontextualised representations: Examining the case of differential calculus through a comparative judgement technique. *Research in Mathematics Education, 22*(3), 284–303. https://doi.org/10.1080/14794802.2019.1692060

Bokhove, C. [@cbokhove]. (2019, June 27). *The time is absolutely not ripe for Comparative Judgement at a national scale in year 6 sats. Validity, *total* time, accountability, logistics are challenges. Maybe just as additional moderation. I feel such desires mainly based on dislike of current system, not eval of CJ.* Twitter. https://twitter.com/cbokhove/status/1144137828621680640?s=43&t=ohKI7Uu4ddr7VZQXzHd2sA

Davis, A. J. [@ded6ajd]. (2017, April 8). *Comparative judgment drains from the assessment any considerations of *what* is being assessed & the whole question of validity worth having.* Twitter. https://twitter.com/ded6ajd/status/850639513399545856?s=43&t=ohKI7Uu4ddr7VZQXzHd2sA

Davies, B., Alcock, L., & Jones, I. (2020). Comparative judgement, proof summaries and proof comprehension. *Educational Studies in Mathematics, 105*(2), 181–197. https://doi.org/10.1007/s10649-020-09984-x

Dolan, T. [@timdolan]. (2021, February 2). *Maths Teachers: Has anyone considered using comparative judgement as part of Y11 or Y13 assessment? I wonder whether it's worth exploring, for data to help with valid inferences about mathematical ability. Would only be norm-referenced data, which may be of limited value?* Twitter. https://twitter.com/timdolan/status/1356532458771087360?s=43&t=ohKI7Uu4ddr7VZQXzHd2sA

Gijsen, M., van Daal, T., Lesterhuis, M., Gijbels, D., & de Maeyer, S. (2021). The complexity of comparative judgments in assessing argumentative writing: An eye tracking study. *Frontiers in Education, 314*. https://doi.org/10.3389/feduc.2020.582800

Heldsinger, S. A., & Humphry, S. M. (2013). Using calibrated exemplars in the teacher-assessment of writing: An empirical study. *Educational Research, 55*(3), 219–235. https://doi.org/10.1080/00131881.2013.825159

Hunsader, P. D., Thompson, D. R., Zorin, B., Mohn, A. L., Zakrzewski, J., Karadeniz, I., Fisher, E. & MacDonald, G. (2014). Assessments accompanying published textbooks: the extent to which mathematical processes are evident. *ZDM, 46*, 797-813. https://doi.org/10.1007/s11858-014-0570-6

Hunter, J., & Jones, I. (2018). Free-response tasks in primary mathematics: A window on students' thinking. *Proceedings of the 41st Annual Conference of the Mathematics Education Research Group of Australasia, 41*, 400–407.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774–1787. https://doi.org/10.1080/03075079.2013.821974

Jones, I., & Davies, B. (2023). Comparative judgement in education research. *International Journal of Research & Method in Education*. https://doi.org/10.1080/1743727X.2023.2242273

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics, 89*(3), 337–355. https://doi.org/10.1007/s10649-015-9607-1

Jones, I., Inglis, M., Gilmore, C., & Hodgen, J. (2013). Measuring conceptual understanding: The case of fractions. In A. M. Lindmeier & A. Heinze (Eds.), *Proceedings of the 37th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 113–120). IGPME.

Jones, I., & Karadeniz, I. (2016). An alternative approach to assessing achievement. In C. Csikos, A. Rausch, & J. Szitanyi (Eds.), *The 40th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 51–58). IGPME.

Jones, I., Swan, M., & Pollitt, A. (2014). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education, 13*(1), 151–177. https://doi.org/10.1007/s10763-013-9497-6

Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation, 47*, 93–101. https://doi.org/10.1016/j.stueduc.2015.09.004

Jones, I., Wheadon, C., Humphries, S., & Inglis, M. (2016). Fifty years of A-level mathematics: Have standards changed? *British Educational Research Journal, 42*(4), 543–560. https://doi.org/10.1002/berj.3224

Kelly, K. T., Richardson, M., & Isaacs, T. (2022). Critiquing the rationales for using comparative judgement: A call for clarity. *Assessment in Education: Principles, Policy & Practice, 29*(6), 674–688. https://doi.org/10.1080/0969594X.2022.2147901

Lok, B., McNaught, C., & Young, K. (2016). Criterion-referenced and norm-referenced assessments: Compatibility and complementarity. Assessment & Evaluation in Higher Education, 41(3), 450–465. https://doi.org/10.1080/02602938.2015.1022136

Marshall, N., Shaw, K., Hunter, J., & Jones, I. (2020). Assessment by comparative judgement: An application to secondary statistics and English in New Zealand. *New Zealand Journal of Educational Studies, 55*(1), 49–71. https://doi.org/10.1007/s40841-020-00163-3

NCTM Research Committee. (2013). New Assessments for New Standards: The Potential Transformation of Mathematics Education and Its Research Implications. Journal for Research in Mathematics Education, 44, 340–352. https://doi.org/10.5951/jresematheduc.44.2.0340

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354

Swan, M., & Burkhardt, H. (2012). Designing assessment of performance in mathematics. *Educational Designer, 2*(5), 1–41. https://isdde.org/wp-content/uploads/2018/05/isdde09_burkhardt_swan.pdf

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–286. https://doi.org/10.1037/h0070288

van Daal, T., Lesterhuis, M., de Maeyer, S., & Bouwer, R. (2022). Validity, reliability and efficiency of comparative judgement to assess student work. *Frontiers in Education, 7*, 1100095. https://doi.org/10.3389/feduc.2022.1100095

Verhavert, S., Bouwer, R., Donche, V., & Maeyer, S. D. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice, 26*(5), 541–562. https://doi.org/10.1080/0969594X.2019.1602027

Wheadon, C., Barmby, P., Christodoulou, D., & Henderson, B. (2020). A comparative judgement approach to the large-scale assessment of primary writing in England. *Assessment in Education: Principles, Policy & Practice*, 27(1), 46–64. https://doi.org/10.1080/0969594X.2019.1700212

The students' use of visual representation for stimulating their metacognitive strategies: Teachers' perspectives

Amal Kadan-Tabaja

Haifa University, Faculty of Education, Israel; math33444@gmail.com

Visual representation is effective in enhancing mathematical learning and thinking processes. This study focuses on visual representations automatically provided by a formative assessment platform to describe students' mathematical strategies. We examined the teachers' perspectives on how the students' metacognitive strategies could be stimulated by visual representation of strategy (VRS) provided by a formative assessment platform in assignments for comparing fractions. Twenty-five teachers participated in this study. Based on different data resources, we were able to identify three categories where the teachers considered the VRS as a tool for stimulating students' use of metacognitive strategies: as part of class management, as part of task requirements, and as part of feedback information.

Keywords: Example-eliciting task, strategy, fraction, visual representation, metacognition.

Introduction and theoretical background

Mathematical strategies refer to the methods used by students to solve problems with mathematical content, whether their answers are correct or not (Hegedus & Otálora, 2022). Researchers have reported that students use different strategies to successfully compare fractions, which serve as the mathematical content of this study. These strategies were included in the mathematical curriculum and taught in the classroom. The common strategies are using the same numerator, the same denominator, the benchmark of one whole or one half, the distance of each fraction from one whole, and finding equivalent fractions by expansion or reduction algorithms to get a common denominator or numerator. Ellis et al. (2019) argued that investigating mathematical features and properties of sets of examples may shed light on the students' strategies and thinking. In this study, we used STEP (Seeing the entire picture) as a formative assessment platform. STEP enables primarily: (a) example-eliciting tasks (EETs)—a task that includes an interactive diagram and asks students to generate examples by performing dragging under given constraints (Yerushalmy & Olsher, 2020); (b) assessing students' mathematical strategies based on automated analysis of the mathematical features of the students' examples; and (c) reflecting these strategies automatically and visually, which we call "visual representation of strategy" (VRS) (Kadan-Tabaja & Yerushalmy, 2023). Research shows that the technological platforms that provide an immediate picture of students' work may support the work of teachers and allow for better representation of the mathematical content and more effective student learning (Olsher, Yerushalmy, & Chazan, 2016). Research shows that students dealing with visual representation related to examples they had constructed may be more effective for their learning and further stimulate their thinking process (Robutti, 2010). In this study, we focused on the students' metacognitive strategies as the thinking process that refers to the learners' knowledge, planning, monitoring, and evaluating their strategies for learning and thinking in the cognitive process (Pintrich, 2002).

The novelty of this study lies in integrating the formative assessment platform with the students' answers to an EET, to represent their mathematical strategies automatically and visually. We examined the teachers' perspectives on how the students used the VRS that STEP provided to stimulate their metacognitive strategies. Our research question was: From the teachers' perspective, how can the students' use of VRS within a formative assessment platform stimulate their metacognitive strategies?

Research setting

The interactive diagram of the task in this study was based on a similar representation as that mentioned in the literature (Figure 1) (Arnon et al., 2001), using the STEP platform. Arnon et al. (2001) studied and reported on students who learned with the Shemesh software, which was designed to promote conceptual learning of fractions, offering concrete representations of the fraction and the operations performed on it. Fractions are represented in the discrete Cartesian coordinate system by a point whose vertical coordinate is the numerator and its horizontal coordinate the denominator. All equivalent fractions are represented on a straight ray passing through the origin point. The origin and points on the vertical axis do not represent any fraction. All equivalent fractions are represented on a straight ray passing through the origin (e.g., $\frac{2}{5}=\frac{4}{10}=\frac{6}{15}$). Points that exist on a ray with a larger slope represent larger fractions (e.g., $\frac{6}{6}>\frac{2}{5}$). The red point (corresponding to an X in Figure 1) represents the given fraction (a fixed point), and the green point (corresponding to an empty circle in Figure 1) is the fraction that the student can drag freely to satisfy the requirement of the task.

The task on which this study was based, students were asked to construct a fraction larger than the given fraction $\frac{2}{5}$ represented by the red point by dragging the green point (Figure 2). The task required students to construct 10 examples of fractions that fit the requirement. Examples were submitted separately and captured in STEP. In response to the students' submissions of their examples, STEP visually represented each student's examples on a single screen using blue points and provided automated analysis both to teachers and students. Automated analysis for teachers represented the students' identified strategies of choosing examples both verbally and visually (Kadan-Tabaja & Yerushalmy, 2023). The automated analysis for students included a set of statements describing strategies of comparing fractions. Students were asked to read and activate each statement, then to mark which statement they used to choose their examples. Otherwise, to indicate the method they used to choose their examples. The statements marked by the students were automatically reflected as VRSs (Figure 3).

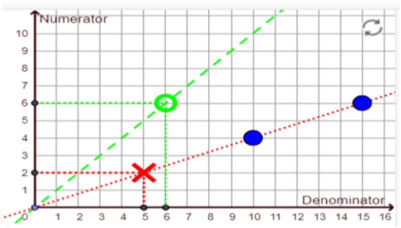


Figure 1: The interactive diagram

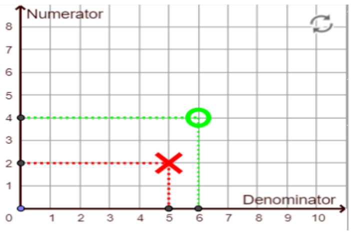


Figure 2: The interactive diagram of the task

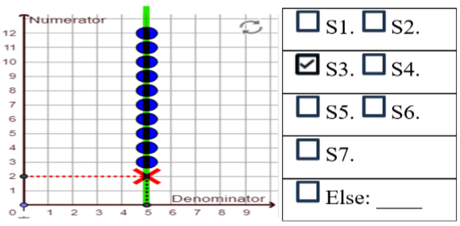
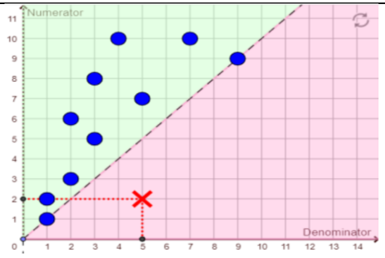
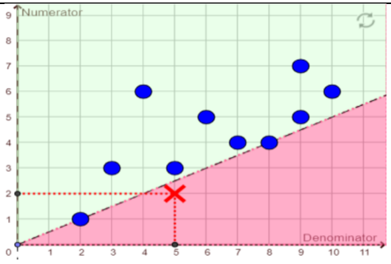
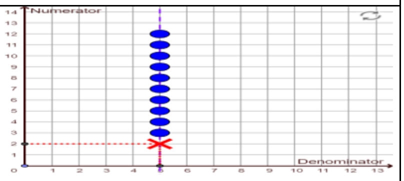
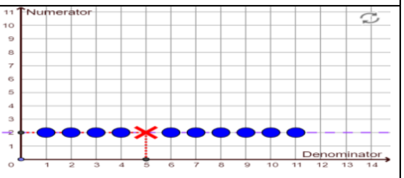
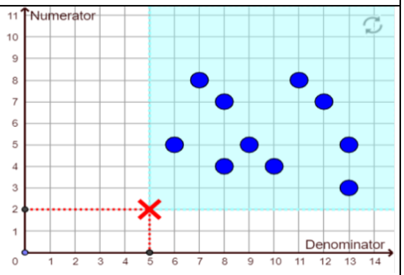
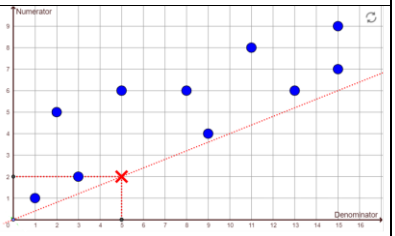
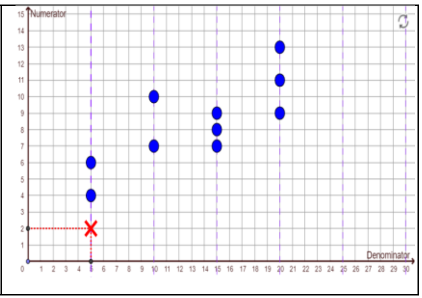


Figure 3: The automated analysis for students

The VRSs are based on automated analysis executed by mathematical algorithms that the researchers set up in STEP when they designed the task. In Table 1, based on Kadan-Tabaja and Yerushalmy (2023), we describe the strategies for comparing fractions (marked in underlined lowercase letters) and the automated VRSs that STEP provides. The teachers were asked to refer to these VRSs when they related to the students' use of the visual representations.

Table 1: Strategies for comparing fractions and the VRSs

The strategies for comparing fractions	**VRS**
S1. Comparing fractions by using a benchmark of one whole. The VRS takes the shape of a ray that separates between two regions. The green/red regions represent all the fractions that are larger/smaller than one, respectively. The ray between them represents fractions equivalent to one whole. The red fraction is less than one, and the submitted examples are equal to or larger than one. Thus, each blue fraction is larger than the red one.	
S2. Comparing fractions by using a benchmark of one-half. The VRS takes the shape of a ray that separates between two regions. The green/red regions represent all fractions that are larger/smaller than one-half, respectively. The ray between them represents fractions equivalent to one-half. The red fraction is less than one-half, and the submitted examples are equal to or larger than one-half. Thus, each blue fraction is larger than the red one.	
S3. Comparing fractions by using the same denominator. The VRS takes the shape of a ray that is parallel with the vertical axes in X= (the denominator of the red fraction). All submitted examples (blue) have the same denominator as the given red fraction.	
S4. Comparing fractions by using the same numerator. The VRS takes the shape of a ray parallel with the horizontal axes in Y= (the numerator of the red fraction). All submitted examples (blue) have the same numerator as the given red fraction.	
S5. Comparing fractions based on "the numerator and the denominator of the larger fraction are larger than the numerator and the denominator of the smaller fraction, respectively." The VRS takes the shape of a region that represents all fractions that have a larger numerator and denominator than the given red fraction. This is a misconception, and each blue fraction can be larger or smaller than the red one.	
S6. The blue fraction is larger than the red one. This strategy is true when the student's example fulfills the requirements of the tasks. In the visual representation, each blue fraction is above the red ray which represents all the fractions that are equivalent to the red fraction, which means that each blue fraction is larger than the red one.	

S7. Comparing fractions based on "the denominator of the larger fraction is a multiple of an integer of the denominator of the smaller fraction." The VRS takes the shape of rays that represent all fractions that are parallel with the vertical axes, X= a × (the denominator of the red fraction), where a is an integer that is not equal to zero. Thus, each blue fraction can be larger than the red one.	

Methodology

Twenty-five elementary and secondary teachers volunteered to participate in the study. The teachers took part in 30 hours of teacher development workshops (10 sessions, 3 hours each) conducted by the first researcher, after which they conducted 4 activities on fractions that contained about 12 tasks. For this study, we focused on one task. Some of the sessions were held online, while others were conducted offline (Table 2).

Table 2: Teacher development workshops

Online/offline sessions	Duration	The process of teachers' works in the session
Online	12 hours	The teachers worked on the tasks in small groups (including the task in our study). They discussed the task requirements, the conceptions or misconceptions, the learning process, the different kinds of feedback, the use of the automated information that the platform enables. Each group documented the discussion in Google Slides and represented it in front of all development workshop participants.
Online	6 hours	The teachers were acquainted with the terms "students' self-regulation," "thinking process," and "metacognitive actions and reflection process."
Offline	12 hours	In the offline sessions of the workshops, each teacher was asked to observe one student engaging with the tasks through the automated VRS that STEP enabled, specifically with the task of this study. Then, the teachers were asked to answer a semi-structured questionnaire containing 10 questions.

Data resources and analysis

To answer the research question, we made use of three resources: (a) video recordings of discussions in small groups (of five teachers each), with the teachers working on the task using the automated VRS in the online session. The transcripts of video recordings were analyzed to extract the statements showing the teachers' perceptions of how their students handle the task and the automated VRSs and how these representations may stimulate the students' metacognitive strategies. (b) While observing the group discussions, the first researcher took field notes. (c) Each teacher responded to a semi-structured questionnaire after observing one student working on the task using the automated VRS. The questionnaire contained ten questions regarding the examples and strategy the student used; challenges; difficulties faced by the student and the way the student handled them; and the insights that the teachers gained from the engagement with the task and the VRS to stimulate the students' thinking and learning. The analysis of the responses to the questionnaire and repeated reading of the data from the transcripts allowed classifying the teachers' perceptions into categories The final categorization of the data was checked by other researchers for consistency (the questions are listed in Table 3).

Table 3: List of questions in the questionnaire the teachers answered after observing one student working on the task using the automated VRSs

1.	Provide background information about the previous knowledge of the student.
2.	Is there a special reason for choosing this student?
3.	What were the characteristics of the fractions that the student chose?
4.	Which strategy did the student choose while constructing the examples?
5.	Were you able to identify difficulties, common mistakes, or misconceptions while the student was working on the task?
6.	Did the student change his/her choices while constructing his/her examples, and if yes, why?
7.	When the student's responses when he/she was exposed to the VRS?
8.	Do you have any suggestions for changes to the current task or the follow-up tasks? Explain.
9.	What are the insights that you gained from the task and the automated assessment? Explain.
10.	In your opinion, how can the use of the VRSs enhance the student's learning and thinking process?

We used a qualitative approach to analyze the data. In an open coding process, we examined excerpts from the teacher's responses and transcripts to describe the categories that would allow us to learn about how, in the teachers' opinion, the use of the VRS stimulated the students' metacognitive strategies. Based on Schoenfeld's (2013) metacognitive framework, we identified the following phrases and sentences that may reflect the metacognitive strategies (Table 4).

Table 4: Metacognitive action and statement

Metacognitive strategy	**When did it occur?**	**The following statement or action may be an example of a metacognitive strategies**
Planning	Before beginning a task	To understand what makes a correct answer… To set specific strategies before beginning a task… To make it easy… To reread the problem…
Monitoring	During the learning and feedback processes	To check the answers and the strategy while working on the task... To ask questions…
Evaluating	After a learning episode	To summarize the learning or thinking after finishing… To evaluate the conclusion that was reached…

Results

The findings show that the teachers' statements according to which the students' use of the VRS may stimulate their metacognitive strategies can be classified into three categories. Below we describe these categories, giving examples from the teachers' statements, and point out the metacognitive strategies that may be stimulated.

a. The VRS is part of classroom management. This category is related to the use of the VRS when students work on the task individually, in pairs, in small groups, or whole-class discussion. For example, one teacher said in the course of an online session: "I think that working on different VRSs in small groups or pairs may help students rethink their answer and compare it with those of others." Another teacher answered in the questionnaire: "It will be very interesting to have a discussion about the different visual representations and to connect them with the strategies for comparing fractions in the classroom." These two examples show that monitoring is a metacognitive strategy that may be stimulated by using the VRSs.

b. The VRS is part of task requirements (or task design), as when using the automated VRS as part of self-reflection before submitting the examples or when the task requirement is specifically related to the visual representation. One teacher stated: "When the student is asked to reflect on the VRS as part of the task requirements, this may be a clue for the student in the choice of examples." Another teacher said in the online session discussion: "We can use the visual representations in tasks to help students rethink and ask questions about their strategies for comparing fractions." According to the first example, planning is the metacognitive strategy that may be stimulated by using the VRS. The second example suggests that monitoring is the metacognitive strategy that may be stimulated by using the VRSs.

c. The VRS is part of feedback information. This category is related to the use of the automated VRS when it is related to the correctness, characteristics, and strategy chosen, and to misconceptions or common mistakes in their examples. For example, one teacher stated that "when the student was exposed to the VRS, he wondered which VRS he might have received had he used another strategy. So, the feedback visual representation made the student think again about the task and about his examples." Another teacher claimed: "The visual representations helped the student check whether his examples were correct or not; he was also able to check the strategy he used for comparing fractions visually." The first example demonstrates that monitoring is the metacognitive strategy that may be stimulated by using the VRSs. The second example demonstrates that evaluating is the metacognitive strategy that may be stimulated by using the VRSs.

Table 5 shows categorization of the metacognitive strategies described by the teachers.

Table 5. Categorization of the metacognitive strategies described by the teachers

The metacognitive strategy	**The visual representation of the strategy is part of class management**	**The visual representation of the strategy is part of task requirements**	**The visual representation of the strategy is part of feedback information**
Planning		The students' use of the VRS may help them understand what makes a correct answer for the task; it may encourage them to look at a broader range of rich strategies and to compare the information across different strategies; and it may help them visualize the sequence of steps of their answer.	

Monitoring	Clustering the students in pairs or small groups for work on tasks using the VRS may help them present, compare, and discuss their strategies with other students. This may enhance the student's mathematical discourse. The students' use of VRSs in the classroom discussion may help them rethink their answers and see the problem from different perspectives.	The students' use of the VRS as part of the task requirement may help them compare, rethink, and adjust their strategy. It may help expose students to strategies and initiate an inquiry process to comply with the task requirements; it may encourage students to generate new strategies based on reasoning and exploration. It may support students in assuming ownership of their learning process and responsibility for it. It may help students enrich their strategic example space. It may encourage students to think in various modes and understand the concept from multiple perspectives.	The students' use of the VRS as part of the feedback information may lead them to assume responsibility for their learning and become independent thinkers. It may help students rethink their examples and ask questions.
Evaluating	The students' use of the VRS in the classroom discussion may expose them to new strategies or interesting answers, which may be used for clarification of their questions or comments.	The students' use of the VRS may help them assemble the lists of evaluation criteria of different strategies by which to assess their examples. It may encourage students to reconsider their answers before submitting them.	The students' use of the VRS as part of feedback information may help them evaluate the correctness and characteristics of their answers, the strategy they used to choose their examples, and the misconceptions or common mistakes in them.

Discussion and conclusion

In this study, we used Schoenfeld's (2013) metacognitive framework to describe, from the teachers' perspectives, how the visual representation of strategy in an automated assessment platform may stimulate the students' metacognitive strategies. Based on the teachers' responses, we were able to identify three categories that related to the students' use of VRSs included in the formative assessment platform and the way each category may stimulate the students' metacognitive strategies—the use of the visual representation when it was part of classroom management, part of task requirements, and part of the feedback information. In each category, teachers stated that monitoring and evaluating actions may stimulate metacognitive strategies. Planning was identified in the students' use of the visual representation when it is part of the task requirements. The findings are consistent with the literature, which reported that visual representation may stimulate the students' mathematical thinking

(Boonen, 2016) and play a central role in the formative assessment process, as perceived by the teachers (Kadan-Tabaja, & Yerushalmy, 2023).

From the teachers' perspective, integrating EETs that automatically and visually reflect students' mathematical thinking into the formative assessment platform appears to stimulate students' metacognitive strategies. The results of this study open new possibilities for using such VRSs in the automated assessment processes in the mathematics classroom.

References

Arnon, I., Nesher, P., & Nirenburg, R. (2001). Where do Fractions Encounter their Equivalents? Can this Encounter Take Place in Elementary-School?. International Journal of Computers for Mathematical Learning, 6(2), 167–214. https://doi.org/10.1023/a:1017998922475

Ellis, A. B., Ozgur, Z., Vinsonhaler, R., Dogan, M. F., Carolan, T., Lockwood, E., ... & Zaslavsky, O. (2019). Student thinking with examples: The criteria-affordances-purposes-strategies framework. The Journal of Mathematical Behavior, 53, 263–283. https://doi.org/10.1016/j.jmathb.2017.06.003

Hegedus, S. J., & Otálora, Y. (2022). Mathematical strategies and emergence of socially mediated metacognition within a multi-touch Dynamic Geometry Environment. Educational Studies in Mathematics, 1–19. https://doi.org/10.1007/s10649-022-10170-4

Kadan-Tabaja, A., & Yerushalmy, M. (2023). The role of automated assessment in representing students' strategies for comparing fractions. In *Thirteenth Congress of the European Society for Research in Mathematics Education (CERME13)*.

Olsher, S., Yerushalmy, M., & Chazan, D. (2016). How might the use of technology in formative assessment support changes in mathematics teaching?. *For the learning of mathematics*, *36*(3), 11-18.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessment. *Theory into Practice, 41*(4), 219-25.

Robutti, O. (2010). Graphic calculators and connectivity software to be a community of mathematics practitioners. *ZDM – Mathematics Education*, *42*, 77-89.

Schoenfeld, A.H. (2013). Reflections on problem solving theory and practice. *The Mathematics Enthusiast, 10*(1–2), 9–34

Yerushalmy, M., & Olsher, S. (2020). Online assessment of students' reasoning when solving example-eliciting tasks: Using conjunction and disjunction to increase the power of examples. ZDM, 52(5), 1033–1049. https://doi.org/10.1007/s11858-020-01134-0

Using e-assessment for interactive example-generation tasks

George Kinnear[1] and Igor' Kontorovich[2]

[1]The University of Edinburgh, UK; G.Kinnear@ed.ac.uk

[2]The University of Auckland, New Zealand; i.kontorovich@auckland.ac.nz

The paper concerns the use of e-assessment systems to advance students' example spaces of mathematical concepts. We report on a pilot study with seven first-year students of linear algebra. Three students engaged with the static version of the e-task, which asked them to generate three examples that were as different as possible. The remaining students were prompted by an interactive e-task that assessed the provided examples and asked for another one with different properties. The analysis of a single task about matrices showed that the students who worked with the interactive e-task were more successful than those working with the static e-task, in terms of the number and the range of generated examples. These preliminary findings open the doors to further study of the design and use of interactive example-generation tasks in university mathematics education.

Keywords: Assessment, concepts, feedback, learner-generated examples.

Introduction

Example-generation tasks have been suggested as an effective way to promote students' learning about concepts (Watson & Mason, 2005), and as a way for researchers (and teachers) to gain insight into students' understanding of concepts (Zazkis & Leikin, 2007). E-assessment offers the potential to provide large undergraduate classes with formative example-generation tasks, giving automated feedback on a scale that would not be feasible for teachers to do manually (Sangwin, 2003). Such formative tasks can serve a dual purpose: on the one hand, students' responses provide the teacher with assessment information about their students' knowledge of concepts; on the other hand, the tasks can also prompt students to consider examples that they might not otherwise think about, thereby promoting further learning about the concepts. The design of e-assessment tasks that expand students' understanding of concepts and the capability to generate concept examples has been identified as an open question that is particularly of interest at the undergraduate level (Kinnear et al., 2022).

Fahlgren and Brunström (2023) note the potential of e-assessment tasks that give feedback in the form of prompts for further examples, with the prompts depending on the examples given so far by the student. Such an approach would automate the recommended approach in clinical interview settings, of "asking for 'another and another' [example] and for 'something different'" (Zazkis & Leikin, 2007, p. 19). Prompting in this way may both stimulate the learners to consider examples beyond the most immediately obvious ones and provide richer data about the learners' knowledge.

Here we describe the design and evaluation of interactive example-generation tasks, that prompt students for further examples based on the examples they have given so far, implemented as prototypes in an e-assessment system. The tasks address topics in linear algebra, and were devised and piloted in collaboration with a group of fourth-year students as part of their undergraduate research project. Our overarching aim was to develop e-assessment tasks that prompt students to generate a rich range of examples. We were particularly interested in the effect of interactivity on students' example-generation activity, so we developed static versions of the tasks to serve as a comparison. The research question guiding this study was: *how do the examples produced by students compare between static and interactive e-assessment tasks?*

Theoretical background

A central theoretical notion of this study is an *example space* (Watson & Mason, 2005), which is a collection of possible examples of a mathematical concept. Watson and Mason define a conventional example space for a given concept as that "generally understood by mathematicians" (p. 62). Each individual has their own personal example space, based on their past experience, and will access it in different ways depending on the situation (e.g., in response to particular cues in a task). The structure of example spaces can be described by the *dimensions of possible variation* (DofPV), which are the features of examples that can vary, and their associated *range of permissible change* (RofPCh). For instance, if asked for a quadratic polynomial, one DofPV is the coefficient of x^2, for which the RofPCh is any non-zero real number. Fahlgren and Brunström (2023) used these notions to analyse three example-generation tasks, and we similarly used them to guide the design and analysis of our tasks.

Zazkis and Leikin (2007) proposed a framework for characterizing students' example spaces, that includes a focus on *accessibility* of the examples: what are the most obvious concept examples, and how readily can students generate examples beyond those. Watson and Mason (2005) offer the metaphor of "example space as larder", where finding an example can be thought of as "either immediately picking out something familiar or having to look for it for a while" (p. 162). According to the *principle of intellectual parsimony*, "when solving a problem, one intends not to make more intellectual effort than the minimum needed" (Koichu, 2010, p. 217), so it may be that students will not consider less accessible examples without explicit prompting.

Watson and Mason (2005) give advice about designing example-generation tasks that prompt students for a range of examples. They suggest asking for a sequence of examples satisfying additional constraints, as the "increasing constraints extend awareness of what is possible" (p. 132). In this way, example-generation tasks can help to draw learners' attention to DofPV (or to the extent of the associated RofPCh) that they were not previously aware of. Arzarello et al. (2011) further argue that helping students to see structure in example spaces through this sort of prompting is a key role for teachers. Our study begins to explore how to prompt students in this way using e-assessment.

Method

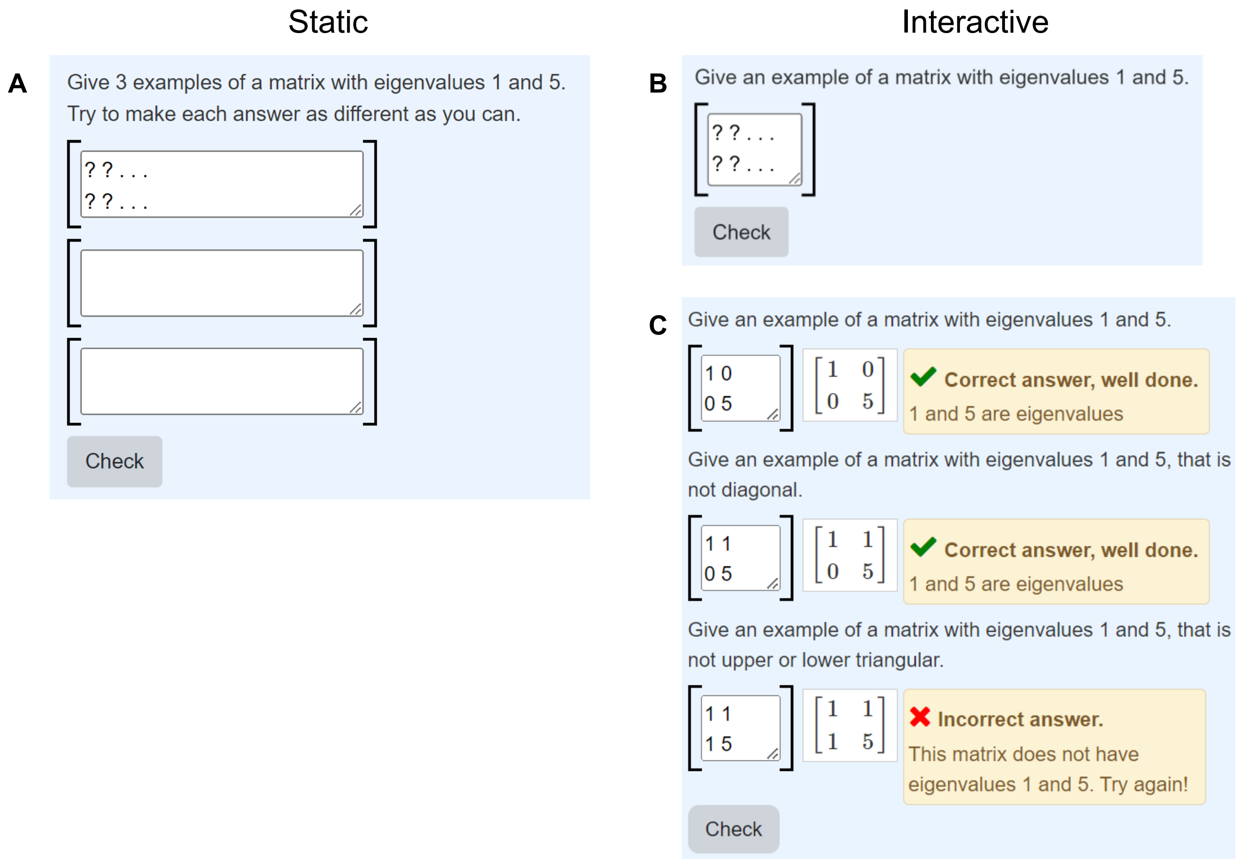
To investigate the potential for interactive example-generation tasks to probe students' example spaces, we developed three prototype tasks for use in clinical interviews with undergraduate students.

Participants and protocol

We invited students taking a first-year linear algebra course to participate in the interviews, near the end of the semester while they were revising for the final assessment. The course is compulsory for students on mathematics and computer science degree programmes, and an option for students on many other degree programmes. The course operates using a flipped classroom design, with preparatory activities that include weekly e-assessment quizzes (for further details about the course, see Docherty, 2023). Out of 581 students on the course, 10 students volunteered to participate in the interviews. During the interview, each student was asked to complete three tasks, with each task in a different format: on paper, as a static e-assessment task, and as an interactive e-assessment task. The task-format combinations were permuted across the 10 interviews, with each task-format combination occurring 3 or 4 times in total. After the students had completed each task, the interviewer asked them to explain how they produced their answers, and whether they could give any other types of examples that had not been covered so far. The interviews were audio-recorded and transcripts were produced for the relevant episodes. However, our main analysis is based on the students' concept examples as recorded by the e-assessment system.

Materials: design of the static and interactive tasks

Three tasks on linear algebra were created for the study, addressing concepts from the course: Eigenvalues, Span, and Reduced row-echelon form. There were two e-assessment versions of each task: static and interactive. The static e-assessment version asked students to provide three examples, and to "try to make each example as different as you can." The interactive e-assessment version asked for a single example; depending on the answers given, students were then prompted for further examples with different properties. Full details for all the tasks can be found at https://osf.io/7wrgz. In this paper, we focus on the Eigenvalues task, which asked students for examples of matrices with eigenvalues 1 and 5 (as shown in Figure 1).



Give 3 examples of a matrix with eigenvalues 1 and 5. Try to make each answer as different as you can.

? ? . . .
? ? . . .

Give an example of a matrix with eigenvalues 1 and 5.

? ? . . .
? ? . . .

Give an example of a matrix with eigenvalues 1 and 5.

1 0
0 5

$\begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$ ✓ Correct answer, well done. 1 and 5 are eigenvalues

Give an example of a matrix with eigenvalues 1 and 5, that is not diagonal.

1 1
0 5

$\begin{bmatrix} 1 & 1 \\ 0 & 5 \end{bmatrix}$ ✓ Correct answer, well done. 1 and 5 are eigenvalues

Give an example of a matrix with eigenvalues 1 and 5, that is not upper or lower triangular.

1 1
1 5

$\begin{bmatrix} 1 & 1 \\ 1 & 5 \end{bmatrix}$ ✗ Incorrect answer. This matrix does not have eigenvalues 1 and 5. Try again!

Figure 1: The Eigenvalues task: the static version (A) asked for three different examples, while the interactive version (B) asked for one initial example before prompting for further examples (C).

For the Eigenvalues task, we anticipated that the most accessible example for students would be the 2 × 2 diagonal matrix with entries 1 and 5 on the diagonal. We designed the interactive version to prompt for further examples by exploring the DofPV that we identified while devising the task. One DofPV is the type of matrix: in addition to diagonal matrices, upper/lower triangular and non-triangular matrices are possible. According to the principle of intellectual parsimony (Koichu, 2010), we expected that students would first try a triangular example (since that would avoid the need for any calculation of eigenvalues) and would only resort to constructing a non-triangular example if specifically prompted to do so. Another DofPV is the size of the matrix, which can be $n \times n$ for any $n \geq 2$ (i.e., the range of permissible change is $n \geq 2$). This dimension may not be immediately obvious to students (despite the course dealing with matrices of different sizes), since the two eigenvalues provided in the task may cue students to think of 2 × 2 examples. Moreover, the task is deliberately vague in not specifying that 1 and 5 are the *only* eigenvalues, and in not specifying their multiplicity.

We implemented the interactive tasks in the STACK e-assessment system, using the feedback messages displayed after a student submitted an answer. If the answer was incorrect, a feedback message was displayed (e.g., "This matrix does not have eigenvalues 1 and 5. Try again!"); the student could then change their answer and re-submit. If the answer was correct, the feedback message included hidden JavaScript code that revealed the next prompt and input box (the code is available at https://osf.io/7wrgz). We opted to use this relatively simple design, even though it imposed a constraint on the decisions about which prompt to show next: these decisions could only be based on the first and last submitted answers, rather than the full sequence of examples provided so far. In line with this constraint, we developed a flowchart for deciding which prompt to present next. The flowchart for the interactive Eigenvalues task is shown in Figure 2.

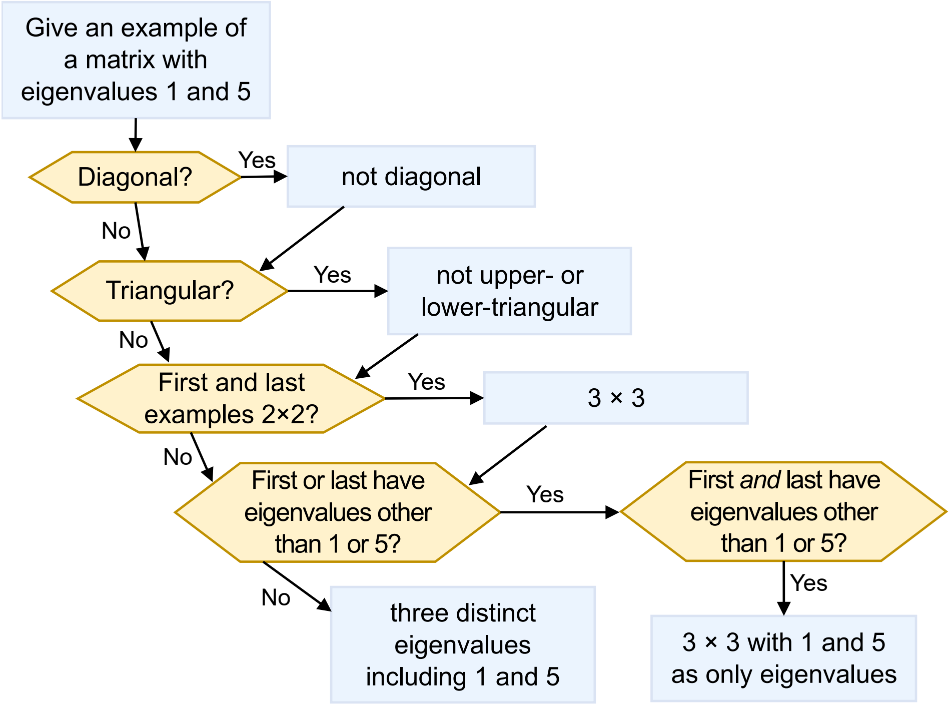

Give an example of a matrix with eigenvalues 1 and 5
Diagonal?
Yes
not diagonal
No
Triangular?
Yes
not upper- or lower-triangular
No
First and last examples 2×2?
Yes
3 × 3
No
First or last have eigenvalues other than 1 or 5?
Yes
First *and* last have eigenvalues other than 1 or 5?
Yes
No
three distinct eigenvalues including 1 and 5
3 × 3 with 1 and 5 as only eigenvalues

Figure 2: Flowchart showing the properties that were checked to determine the next prompt in the interactive Eigenvalues task

Results

The student responses to the Eigenvalues task are shown in Table 1 (interactive version; 3 students) and Table 2 (static version; 4 students). All students gave the diagonal matrix with entries 1 and 5 as their first example, confirming our expectation that this would be the most accessible example.

For the interactive version of the Eigenvalues task, after the initial diagonal example, only one of the three students (S1) proceeded to produce a triangular example as expected (i.e., the triangular examples were less accessible than we had anticipated). The other two students (S2 and S10) moved immediately to considering the characteristic polynomial of a general 2 × 2 matrix and produced non-triangular examples. S2's example was incorrect; the student was stuck at this point so the interviewer intervened with a correct example so they could proceed to the next prompt. S1 similarly produced an incorrect 2 × 2 example when the task prompted them for a non-triangular matrix, however they switched (unprompted) at that point to consider 3 × 3 examples. When asked to explain their method after completing the task, S1 said that "there probably is a way" to make a 2 × 2 non-triangular example. All three students were able to produce 3 × 3 examples with the required properties.

Table 1: Student responses to the interactive Eigenvalues task.

	Student 1	Student 2	Student 10
	$\begin{bmatrix}1&0\\0&5\end{bmatrix}$	$\begin{bmatrix}1&0\\0&5\end{bmatrix}$	$\begin{bmatrix}1&0\\0&5\end{bmatrix}$
Prompt 1	*Not diagonal?*	*Not diagonal?*	*Not diagonal?*
	$\begin{bmatrix}1&1\\0&5\end{bmatrix}$	$\begin{bmatrix}\sqrt{2}&\sqrt{2}\\-5\sqrt{2}&5\sqrt{2}\end{bmatrix}$ (✗)	$\begin{bmatrix}3&2\\2&3\end{bmatrix}$
		[interviewer intervened]	
Prompt 2	*Not triangular?*	*3×3?*	*3×3?*
	[incorrect 2×2] $\begin{bmatrix}1&0&0\\1&1&1\\0&0&5\end{bmatrix}$	$\begin{bmatrix}1&0&0\\0&5&0\\0&0&1\end{bmatrix}$	$\begin{bmatrix}1&0&0\\0&1&0\\0&0&5\end{bmatrix}$
Prompt 3	*Three distinct eigenvalues?*	*Three distinct eigenvalues?*	*Three distinct eigenvalues?*
	$\begin{bmatrix}1&0&0\\0&5&0\\0&0&2\end{bmatrix}$	$\begin{bmatrix}1&0&0\\0&5&0\\0&0&3\end{bmatrix}$	$\begin{bmatrix}1&0&0\\0&5&0\\0&0&2\end{bmatrix}$

Four other students attempted the static Eigenvalues task (see Table 2). Two of the students (S5, S9) only managed to provide the diagonal example; while both students spent several minutes working on paper with the characteristic polynomial of a general 2 × 2 matrix, neither was able to generate a further non-diagonal example, and neither considered larger matrices (although, when prompted at the end of the interview, both students were able to produce further examples). The other two students (S6, S7) generated further examples after the diagonal matrix. S7 noted that the task "didn't say anything about the multiplicity" when explaining their use of a 3 × 3 matrix for the second example. They also demonstrated awareness of the generality represented by this example (an upper-triangular matrix) when explaining their answer: "If it is a triangular matrix then the determinant is the product of the leading diagonal. If you take the characteristic polynomial it ends up $(x-1)(x-1)(x-5)$”. They were not confident in their final example (“it’s possibly wrong”) but reasoned using the cofactor expansion for computing determinants, where “I think the zeros would cancel out.” S6 also reasoned using the cofactor expansion, but gave an incorrect 3 × 3 example. When asked how they produced this example, they explained that when choosing the matrix entries, “as long as I have one row or one column of zeros, I can fill in any random values” (i.e., they appeared to overlook the need to consider the determinant of the bottom-right 2 × 2 sub-matrix). S6 was able to produce a correct non-diagonal example: an upper-triangular 2 × 2 matrix that they described as “a variation of” their first example.

Table 2: Student responses to the static Eigenvalues task.

Student 5	**Student 6**	**Student 7**	**Student 9**
$\begin{bmatrix}1 & 0\\0 & 5\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 5\end{bmatrix}$ $\begin{bmatrix}1 & 0 & 1\\0 & 5 & 2\\0 & 3 & 1\end{bmatrix}$ (✗) $\begin{bmatrix}1 & 2\\0 & 5\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 5\end{bmatrix}$ $\begin{bmatrix}1 & 3 & 6\\0 & 1 & 4\\0 & 0 & 5\end{bmatrix}$ $\begin{bmatrix}5 & 0 & 0\\2 & 1 & 9\\5 & 0 & 5\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 5\end{bmatrix}$

A summary of the results is shown in Table 3, where each student’s examples are classified according to the two DofPV that we identified for this task: the type of matrix and size of matrix. This enables comparison between students (and between the two task modalities), in terms of which DofPV they showed awareness of through their examples (e.g., providing both 2 × 2 and 3 × 3 examples shows awareness of the “size of matrix” DofPV). For the static version, S5 and S9 did not show awareness of either DofPV since they provided only one example, while S6 and S7 showed awareness of both DofPV. For the interactive version, all three students showed awareness of both DofPV, although only S1 provided a triangular example within the “type of matrix” DofPV.

Table 3: Summary of student responses to the Eigenvalues task showing awareness of DofPV as evidenced by correct (filled disc) or incorrect (hollow disc) examples.

		Type of matrix			Size of matrix	
		Diagonal	Triangular	General	2 × 2	3 × 3
Static	S5	●			●	
	S9	●			●	
	S6	●	●	○	●	○
	S7	●	●	●	●	●
Interactive	S2	●		○	●	●
	S10	●		●	●	●
	S1	●	●	●	●	●

Discussion

We developed prototypes of interactive example-generation tasks and tested them with students in a laboratory setting, alongside static versions of the same tasks. The analysis presented here focused on one task, about matrices with given eigenvalues. Overall, the interactive version of the task appears to have been successful in prompting students to consider both DofPV that we had targeted in the design of the task. This stands in contrast to the static version of the task, where two of the four students provided only a single example and therefore did not demonstrate awareness of any DofPV (though they were later able to do this with explicit prompting from the interviewer).

Interactive example-generation tasks have the advantage of providing a controlled approach for prompting learners to generate examples, which is important when seeking to "make inferences about participants' knowledge from the examples they generate" (Zazkis & Leikin, 2007, p. 19). Since all the prompts in the e-assessment task need to be decided in advance, careful thought should be given to the sequence of prompts; and where the e-assessment system has powerful mathematical capabilities, properties can be checked more quickly and reliably than an interviewer or a teacher could manage to do on the spot. However, the pre-designed prompts may turn out to be sub-optimal when learners respond in unanticipated ways. For instance, with the Eigenvalues task, S2 and S10 were not prompted to produce a triangular example since they skipped over that step in the anticipated sequence of "diagonal → triangular → non-triangular" (i.e., the triangular examples were less accessible than was anticipated). This demonstrates the value of pilot studies in the development of interactive example-generation tasks, so that the design can be refined in light of students' responses.

The results of this pilot study of the Eigenvalues task suggest that the flowchart that we developed (Figure 2) could be refined to ensure that students are prompted to consider the full range of anticipated DofPV. We have already begun to work on a more sophisticated implementation, that could make decisions about which prompt to give based on details of the entire sequence of examples generated so far: for instance, "both of your examples are 2 × 2, can you give one that is 3 × 3?", or "could you think of a simpler example, like a triangular matrix?". This approach could be combined with asking for two or three different examples at the outset (similar to Fahlgren & Brunström, 2023), to get some indication of the most accessible examples and the DofPV that the student is aware of, and thus determine which DofPV would be fruitful ones to explore next.

Two further improvements could be considered for this task, which may also apply to other tasks. First, it could be worthwhile to offer students hints about how to proceed if they struggle to generate an example, even for types of example that are expected to be the most accessible. For instance, S2 was unsure how to proceed after their second example was incorrect; the interviewer intervened to move them on, and without this intervention they may not have had the chance to consider other DofPV. This sort of intervention could perhaps be formalized, by giving the student the option of asking for a hint. Second, the task could include prompts for examples that are not possible, to probe the students' understanding of the RofPCh for particular DofPV. For instance, students could be asked to "give an example of a 2 × 2 matrix with eigenvalues 1, 2 and 3 (or enter none if this is not possible)".

This was a small-scale pilot study of these tasks, so we do not seek to make any strong claims about the comparison between the interactive and static task formats. However, our findings do suggest that the interactive example-generation task was able to stimulate learners to consider a broad range of examples, and that this range was broader than the one demonstrated by the static task group. Indeed, students completing the static versions of the tasks admitted to giving up on the instruction to produce examples that were “as different as possible”. We believe that the interactive example-generation task format therefore warrants further development and study.

Acknowledgment

We would like to thank Molly Berrick, Abbie Whitehouse, Eilidh Millin and Álvaro Beattie Eizaguirre who collected data for this study as part of their fourth-year project at the University of Edinburgh.

References

Arzarello, F., Ascari, M., & Sabena, C. (2011). A model for developing students’ example space: The key role of the teacher. *ZDM*, *43*(2), 295–306. https://doi.org/10.1007/s11858-011-0312-y

Docherty, P. (2023). Case study 3: An introductory linear algebra course. In A. Wood (Ed.), *Effective Teaching in Large STEM Classes* (pp. 9-1–9-9). IOP Publishing. https://doi.org/10.1088/978-0-7503-5231-4ch9

Fahlgren, M., & Brunström, M. (2023). Designing example-generating tasks for a technology-rich mathematical environment. *International Journal of Mathematical Education in Science and Technology*, 1–17. https://doi.org/10.1080/0020739X.2023.2255188

Kinnear, G., Jones, I., Sangwin, C., Alarfaj, M., Davies, B., Fearn, S., Foster, C., Heck, A., Henderson, K., Hunt, T., Iannone, P., Kontorovich, I., Larson, N., Lowe, T., Meyer, J. C., O’Shea, A., Rowlett, P., Sikurajapathi, I., & Wong, T. (2022). A collaboratively-derived research agenda for E-assessment in undergraduate mathematics. *International Journal of Research in Undergraduate Mathematics Education*. https://doi.org/10.1007/s40753-022-00189-6

Koichu, B. (2010). On the relationships between (relatively) advanced mathematical knowledge and (relatively) advanced problem-solving behaviours. *International Journal of Mathematical Education in Science and Technology*, *41*(2), 257–275. https://doi.org/10.1080/00207390903399653

Sangwin, C. J. (2003). New opportunities for encouraging higher level mathematical learning by creative use of emerging computer aided assessment. *International Journal of Mathematical Education in Science and Technology*, *34*(6), 813–829. https://doi.org/10.1080/00207390310001595474

Watson, A., & Mason, J. (2005). *Mathematics as a Constructive Activity*. Routledge. https://doi.org/10.4324/9781410613714

Zazkis, R., & Leikin, R. (2007). Generating examples: From pedagogical tool to a research tool. *For the Learning of Mathematics*, *27*(2), 15–21.

Leveraging formative assessment to develop students' mathematical reasoning through images and student generated language

Richard Kitchen[1] and Janet Lear[2]

[1]University of Wyoming, School of Teacher Education, Laramie, WY, USA; rkitchen@uwyo.edu

[2]University of Wyoming, School of Teacher Education, Laramie, WY, USA; jlear2@uwyo.edu

A sixth-grade teacher who taught at a highly diverse school in the United States participated in professional development activities during the 2022-23 school year to learn about and implement an instructional protocol designed for use during problem-solving lessons. The protocol is intended to guide teachers to simultaneously attend to developing their students' mathematical reasoning and learning of the mathematics register. In a case study undertaken to examine teachers' use of the protocol, an example emerged of how formative assessment can be leveraged to support the development of students' mathematical reasoning through student generated language and images. This study contributes to the research literature by demonstrating how the use of formative assessment can promote the use of language to "carry" a mathematical concept in the sense that language can help students create a mental image of that concept.

Keywords: Formative assessment, instructional innovation, mathematical reasoning.

Introduction

During the 2022-23 school year, the lead author collaborated with "Ms. Diaz," a sixth-grade teacher who taught at a diverse school district in northern New Mexico, USA. Ms. Diaz volunteered to learn about and use an instructional protocol designed specifically for use during problem-solving lessons. The protocol, referred to as the "Discursive Mathematics Protocol" (DMP), is intended to be used as a guide by teachers to support students to develop both their mathematical reasoning and learn the mathematics register. Additional information about the DMP can be found in Matute (2022). In this paper, a vignette is shared from one of the problem-solving lessons in which Ms. Diaz implemented the DMP. The vignette demonstrates how formative assessment can be leveraged to support the development of students' mathematical reasoning through student generated language and images. The following research question is addressed in this study: What is an example of how formative assessment can used to support students' mathematical reasoning through student generated language? Brief reviews of the research literature on formative assessment and the role of language and images in the learning of mathematical concepts are now provided.

Formative assessment in mathematics

Classroom assessments are used to inform teachers, students and parents about student knowledge and understanding of mathematical concepts, processes and skills (Wiggins, 1993). There are two categories of classroom assessments; summative and formative. Summative assessment formats focus on what students know at a given time (Guskey & Bailey, 2001). Formative assessments differ from summative assessment in that the focus is not just on summarizing students' learning, but on using student learning data to inform instruction. After examining 250 research studies on classroom assessments, Black and Wiliam (1998) found that when teachers focus on formative assessment, student achievement gains are among the largest ever reported for educational interventions. Formative assessment can include any of the following: classroom observation, inquiry, group work, whole class discussions, peer assessment, written work, individual interviews, student self-assessment, and portfolio assessment (Gearhart & Saxe, 2004). The vignette provided below resulted from the interplay of classroom observation, inquiry, group work, and a whole class discussion.

The role of language and images in the learning of mathematics

Words are referents to mental images (Arnheim, 1969) that teachers can leverage to support their students' mathematical reasoning and learning (Gonzales, 2004). As one example, Kieren (1988) described how students used visual images associated with cutting, symmetry, and numerical halving to make sense of and express their ideas about the notion of something being a half of a whole. The deliberate use of images and language is particularly relevant in problem-solving based lessons given that representations are used to understand problems and devise solutions to those problems (Pólya, 1945/1986). In addition, students engage in language-rich activities to problem-solve by making conjectures, conceiving arguments, and formulating and carrying out proofs (Schoenfeld, 2013). The study's theoretical perspective is introduced next followed by an explanation of the research methods.

Theoretical framework

Given the considerable attention given to cognition, sense making, and social discourse in the DMP, our theoretical framework is social-constructivism (Shepard, 2000). In this study, we viewed cognition through a measurement lens as students solve a task introduced below, examined students' mathematical reasoning as they made sense of images and used language to devise solution strategies and test those strategies (Shepard, 2000), and situated student learning as occurring through their engagement in discourse in communities of practice (Erath et al., 2021).

Methods and data sources

This study was part of a case study research project that examined the DMP with teachers such as Ms. Diaz. In case studies, a real-life, bounded system or entity is selected to study within a particular setting (Yin, 2014). Throughout the 2022-23 school year, the following data was collected: observation data, videos of teachers and students who provided consent, student work and interviews with the two participating teachers. The data presented here were collected during one problem-solving lesson led by Ms. Diaz with her sixth-graders. A co-teaching approach (Cook & Friend, 1995) was used to plan and deliver the task. During the implementation of the task "Height Requirements" (See Figure 1), the lead author facilitated whole class discourse with Ms. Diaz as shown below.

At Sea World San Diego, kids are only allowed into the Air Bounce if they are between 37 and 61 inches tall. They are only allowed on the Tide Pool Climb if they are 39 inches tall or under: 1. Represent the height requirements of each ride in words.	3. Show the allowable heights for the rides on separate number lines. 4. Using inequalities and a number line, describe the height of kids who can go on both the Air Bounce and the Tide Pool Climb. Explain how you figured this out.

2. Represent the height requirements of each ride with inequalities. Explain the meaning of the terms you used for each of your inequalities.	
Adapted from IM: https://tasks.illustrativemathematics.org/content-standards/6/EE/B/8/tasks/2010	

Figure 1: Height Requirements Task

We used social-constructivism (Shepard, 2000) to examine students' mathematical reasoning in conjunction with how language was generated by Ms. Diaz and her students to communicate ideas. As will be shown, a vignette emerged as students were attempting to address the task's fourth question which involved determining the range of kids' heights that would allow them to go on the two rides.

Results

In a video vignette described below, the phrase "in between" became a focus of instruction. What inspired the use of this phrase was that "Elena" used it when describing the shaded portion of her number line between 37 and 39 in a small group discussion with her peers and the first author. She shared, "It could just go in between 37 and 39." After a short discussion in which Elena explained further, another girl in her small group explained why the answer must be in between 37 and 39: "Because 37 is how far you have to be for the Air Bounce and 39 is how tall you have to be for the Tide Pool." In a whole class discussion, students in the class were asked to revoice Elena's idea. The following discussion ensued between two students and the lead author:

1 S1: What she's basically saying is, like, you know how the Tide Pool ends at 39 and the Air Bounce starts at 37?
2 RK: Yeah.
3 S1: You say 37, 38, and 39 inches could ride both the rides.
4 RK: Kids who are 37, 38 or 39 inches tall?
5 S1: Yeah.
6 RK: What if the kid is 37 and a half inches tall?
7 S2: They can still ride it.
8 RK: Why? Are they still in between 37 and 39 inches tall?
9 S1: Yeah.

At this point, Ms. Diaz encouraged "Luna" to come forward to share her ideas at the front board. Luna described her double number line answer that shows an overlap of heights that would allow children to take both rides. Luna then proceeded to modify her written response to the task's fourth question displayed on the board by including the phrase "in between" in her response, observed by the entire class. Others leveraged this phrase as well in their written solutions to the task's fourth question, helping them derive the range of kids' heights that would allow them to go on both rides.

Discussion

The use of formative assessment, specifically an observation made of student generated ideas, provided the means to engage students in a whole class discussion on the language they produced to solve the task. In the given vignette, a student's use of the phrase "in between" led to students making sense of the meaning of the overlap of the two number lines. The overlap of two shaded sections of a number line allowed students a visual means to derive the solution to the task. The vignette shows how words are referents to mental images (Arnheim, 1969) and how teachers can leverage both images and language to help students learn mathematics (Gonzales, 2004). Moreover, the vignette reveals how language can "carry" a mathematical concept in the sense that language can help students create a mental image of that concept. Leveraging a simple phrase such as "in between" demonstrates how student generated language can support mathematical understanding. This study contributes to the research literature by giving an example of how the use of formative assessment can promote the use of language to "carry" a mathematical concept to help students create a mental image of concepts.

References

Arnheim, R. (1969). *Visual thinking*. Berkeley, CA: University of California Press.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139-148.

Cook, L., & Friend, M. (1995). Co-teaching: Guidelines for creating effective practices. *Focus on Exceptional Children*, *28*(3), 1-16.

Erath, K., Ingram, J., Moschkovich, J., & Prediger, S. (2021). Designing and enacting instruction that enhances language for mathematics learning: A review of the state of development and research. *ZDM – The International Journal on Mathematics Education*, *53*, 245–262. https://doi.org/10.1007/s11858-020-01213-2

Gearhart, M. & Saxe, G. B. (2004). When teachers know what students know: Integrating mathematics assessment. *Theory Into Practice*, *43*(4), 304–313.

Gonzales, L. (2004). *Building personal knowledge of rational numbers through mental images, concepts, language, facts, and procedures* (Publication No. 3155976) [Doctoral dissertation, New Mexico State University]. ProQuest Dissertations & Theses Global.

Guskey, T. R., & Bailey, J. M. (2001). *Developing grading and reporting systems for student learning*. Corwin Press.

Kieren, T. E. (1988). *Personal knowledge of rational numbers: Its intuitive and formal development. Number concepts and operations in the middle grades* (Vol. 2). Reston, VA: National Council of Teachers of Mathematics.

Matute, K. (2022). *Discursive Mathematics Protocol for supporting English learner students to develop the mathematics register*: *A case study* (Order No. 29321247). Available from Dissertations & Theses @ University of Wyoming. (2718164761).

Pólya, G. (1986). *How to solve it: A new aspect of mathematical method*. Princeton, NJ: University Press.

Schoenfeld, A. H. (2013). Reflections on problem solving theory and practice. *The Mathematics Enthusiast*, *10*,(1–2), 9–34.

Shepard, L.A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29(7)*, 4–14.

Wiggins, G.P. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco, CA: Jossey-Bass.

Unravelling (Mis)conceptions about Algebraic Letters: Exploring Response Patterns in the 'Meaning of Letters' SMART-test using Latent Class Analysis

Katrin Klingbeil[1] and Filip Moons[2]

[1]University of Duisburg-Essen, Germany; katrin.klingbeil@uni-due.de

[2]Freudenthal Institute, Utrecht University, The Netherlands; f.moons@uu.nl

Identifying typical hurdles, common errors and misconceptions in a certain domain is crucial to deepen our understanding of students' learning. In this paper, we explore response patterns of 2051 German grade 7 and 8 students shown in the SMART-test "Meaning of Letters", designed to assess (mis)conceptions regarding variables, more precisely, the letter-as-object misconception. Using Latent Class Analysis, we were able to identify six response patterns. These patterns are described and thoroughly analysed. They urge us to think more deeply about the interplay between (mis)conceptions and contexts and can help build valid assessment tools to diagnose students' current understanding.

Keywords: Online formative assessment, algebra, variables, letter-as-object misconception, latent class analysis.

Introduction

Assessing students' (mis)conceptions is a challenging task. SMART ("Specific Mathematics Assessments that Reveal Thinking") online tests, that have been developed at the University of Melbourne since 2008, offer a solution by facilitating easy provision and processing of diagnostic tasks on students' conceptual understanding and potential misconceptions. SMART's extended analysis detects patterns between diagnostic tasks, revealing insights into students' understanding and misconceptions. In addition to this automatic diagnosis, it also provides teachers with explanations, tasks, and suggestions for targeted interventions (Steinle et al., 2009).

The test investigated here, *Meaning of Letters*, aims to assess the *letter-as-object* misconception and its subtypes based on students' responses to six multiple-choice tasks. Despite known challenges of multiple-choice tasks, developers argue that well-designed tasks can effectively unveil students' thinking: Klingbeil et al. (2024) showed that students' explanations aligned well with their shown (mis)understandings in their multiple-choice responses.

In a comprehensive intervention study spanning six federal states in Germany, 2051 7th- and 8th-grade students undertook the *Meaning of Letters* test after a few algebra lessons. This paper investigates the response patterns of these students using *Latent Class Analysis* (LCA) (Brandenburger & Schwichow, 2023). LCA is a statistical method used to identify unobservable subgroups (latent classes) within a heterogeneous population based on patterns of responses. This analysis can unravel the (mis)conceptions in understanding algebraic letters and how they interact.

In the following paragraphs, we introduce the theoretical background behind the *Meaning of Letters*-test and the six tasks; next we pose the research question.

Struggling to understand algebraic letters: the *letter-as-object* misconception and its subtypes

Arcavi, Drijvers and Stacey (2017) "distinguish five facets of the concept of variable: a placeholder for a number, an unknown number, a varying quantity, a generalised number, and a parameter" (p. 12). Across these facets, variables stand for or refer to one or more numerical values. Yet, algebra learners often struggle with this numerical interpretation, and various typical errors and misconceptions have been identified. One of them, the *letter-as-object* (LO) misconception, has been described by Küchemann in 1981 as the letter being "regarded as a shorthand for an object or as an object in its own right" (p. 104) and extensively documented over decades (e.g., Akhtar & Steinle, 2017). As part of the foundational Concepts in Secondary Mathematics and Science (CSMS) study on the mathematical understanding of secondary school students in the United Kingdom, Küchemann (1981) utilised the following task: "Blue pencils cost 5 pence each, and red pencils cost 6 pence each. I buy some blue and some red pencils and altogether it costs me 90 pence. If b is the number of blue pencils bought and if r is the number of red pencils bought, what can you write down about b and r?" (p. 107).

While only 10% of tested 14-year-old students provided the correct equation $5b + 6r = 90$, 17% gave $b + r = 90$ as an answer, which might have been read as "blue pencils and red pencils together cost 90 pence" (LO). Interpreting b as "the number of blue pencils" is a possibility here, too; however, this would still imply a wrong understanding of equations with a number of pencils on one side of the equation and the price of all pencils on the other. Interestingly, 6% of the students came up with another kind of equation: $6b + 10r = 90$ or $12b + 5r = 90$. These students had figured out a possible solution to the problem first and then used these values as coefficients in their equation. Since the letters are used as abbreviations for the involved objects ("12 blue pencils and 5 red pencils together cost 90 pence"), this is regarded as a special form of LO, which we will refer to as the s*olution-as-coefficient* (SAC) misconception in the following. Another special form of LO is called *letter-as-unit* (LU) when the algebraic letter is interpreted as an abbreviation for a unit (Akhtar & Steinle, 2017), e.g., in a task about 8 trucks weighing 24 tonnes, the t in the equation $8t = 24$ would be misinterpreted as standing for tonnes (not realising that this would not be a correct equality).

Research question

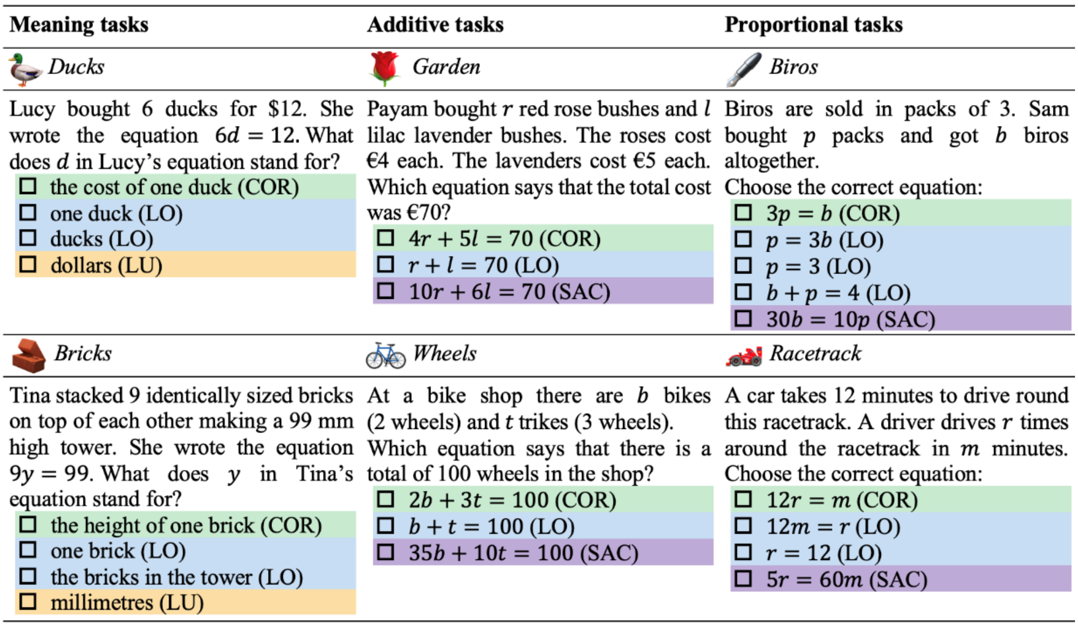
The *Meaning of Letters* SMART-test polls the understanding of variables and detects the presence of the letter-as-object misconception, leading to the following research question: *Which response patterns regarding the letter-as-object misconception can be identified among German grade 7 and 8 students based on their responses to the six multiple-choice tasks of the SMART-test Meaning of Letters?*

Methods and Materials

SMART-test *Meaning of Letters*

For the diagnosis of students, two parallel versions of the SMART-test *Meaning of Letters* were used with the A or B version randomised by class. Here, we describe only the A version of the German translation of the test (see Figure 1; COR indicating the correct response option). The first task type (Meaning tasks), originating from the work of MacGregor and Stacey (1997), uses only one algebraic letter and asks students to decide on the meaning of this letter in a linear equation in a given context.

While the *Ducks* item uses the initial letter of the involved objects and units, the *Bricks* item uses the letter y. Apart from the correct response (cost/height), MC options include the involved objects (singular and plural; LO) as well as the corresponding unit (LU).



Meaning tasks	Additive tasks	Proportional tasks
Ducks	*Garden*	*Biros*
Lucy bought 6 ducks for $12. She wrote the equation $6d = 12$. What does d in Lucy's equation stand for? □ the cost of one duck (COR) □ one duck (LO) □ ducks (LO) □ dollars (LU)	Payam bought r red rose bushes and l lilac lavender bushes. The roses cost €4 each. The lavenders cost €5 each. Which equation says that the total cost was €70? □ $4r + 5l = 70$ (COR) □ $r + l = 70$ (LO) □ $10r + 6l = 70$ (SAC)	Biros are sold in packs of 3. Sam bought p packs and got b biros altogether. Choose the correct equation: □ $3p = b$ (COR) □ $p = 3b$ (LO) □ $p = 3$ (LO) □ $b + p = 4$ (LO) □ $30b = 10p$ (SAC)
Bricks	*Wheels*	*Racetrack*
Tina stacked 9 identically sized bricks on top of each other making a 99 mm high tower. She wrote the equation $9y = 99$. What does y in Tina's equation stand for? □ the height of one brick (COR) □ one brick (LO) □ the bricks in the tower (LO) □ millimetres (LU)	At a bike shop there are b bikes (2 wheels) and t trikes (3 wheels). Which equation says that there is a total of 100 wheels in the shop? □ $2b + 3t = 100$ (COR) □ $b + t = 100$ (LO) □ $35b + 10t = 100$ (SAC)	A car takes 12 minutes to drive round this racetrack. A driver drives r times around the racetrack in m minutes. Choose the correct equation: □ $12r = m$ (COR) □ $12m = r$ (LO) □ $r = 12$ (LO) □ $5r = 60m$ (SAC)

Figure 1: Tasks of Meaning of Letters test (German A version) translated back into English

The second type of task (Additive tasks) is based on Küchemann (1981). It uses two algebraic letters (corresponding to the initial letters of involved objects), which are additively connected and restricted by the given situation. Students are supposed to choose the correct linear equation (in standard form) for the described context. In the correct equation, the letters represent the number of objects and the coefficients for the price per object (*Garden*) and the number of components per object (*Wheels*), respectively. The first alternative response option simply adds the variables without any coefficients, making it possible to interpret the letters as abbreviations for the involved objects (e.g., "Bikes and trikes have 100 wheels altogether."; LO). In the equation of the other alternative option, coefficients equal a possible solution to the problem (that has not been posed) so that the equation can be read as some solution sentence (e.g., "35 bikes (with 2 wheels each) plus 10 trikes (with 3 wheels each) have 100 wheels altogether."; SAC). Also, in this case, the letters are read as abbreviations for the involved objects.

The third task type (Proportional tasks) is derived from the famous "Students and Professors" problem (Clement et al., 1981):

> *"There are six times as many students as professors at this university." Write an equation using S for the number of students and P for the number of professors.*

This is often answered with $6S = P$ instead of $6P = S$. The proportional tasks in the SMART-test have the same algebraic structure: the two variables are directly proportional to each other. Students are again asked to choose the equation matching the given situation. In the correct equation, the letters (matching the initial letters of involved objects/units) stand for the number of objects for both involved objects (*Biros*) or for the number of racetrack rounds and the number of minutes (*Racetrack*). In both items, the coefficient is the proportionality constant (number of biros per pack or number of minutes per round). The first alternative response option is the reverse of the correct equation, which allows for a LO interpretation (e.g., “A pack contains 3 biros.” or “1 round equals 12 minutes”; LO). For the *Racetrack* task, the first alternative can also be seen as an LU interpretation (e.g., “12 minutes equals one round”). However, since it is unclear how exactly students interpret the letter here, we opted for the more general LO interpretation. The LO interpretation also applies to the second alternative response although the second variable is missing (e.g., as “A pack has 3.”; LO). In the equation of the third alternative option, the coefficients correspond to a possible solution (to the question that has not been asked), which can be interpreted as a kind of solution sentence (e.g., “Sam bought 10 packs and has 30 biros now.”), indicating the SAC misconception. The *Biros* task offers one more response option that features the addition of the two variables without coefficients. Again, the letters can be interpreted as abbreviations (e.g., as “One biro plus a pack of biros is 4 altogether.”; LO). This response type does not make sense for the *Racetrack* task since no different objects but rounds and minutes would be added.

Participants

In total, 2051 grade 7 and 8 students (aged 12–14) from six federal states of Germany (78% attending grammar schools, 22% attending non-grammar schools) completed the SMART test. These students were taught by 103 mathematics teachers, leading to a nested data structure (data are analysed at the student level, but the students are clustered in classes).

Teachers were asked to administer the SMART online test 1–2 weeks into their teaching sequence about variables, algebraic expressions and/or equations. Thus, students in grade 7 should have been familiar with a basic concept of variables and be able to use and manipulate them in easy algebraic expressions before taking the test. In grade 8, students probably have started focussing on (solving) equations.

Data analysis

The response patterns of the test were analysed using Latent Class Analysis (LCA) (Brandenburger & Schwichow, 2023). LCA is a form of structural equation modelling useful for identifying patterns/groups within categorical responses. These patterns/groups are called *latent classes.* Intuitively, one can think of the 2051 participating students as ‘latent classes’. Of course, a description of 2051 ‘latent classes’ will be hard to interpret. LCA considerably reduces the complexity of the data by grouping students with similar patterns of responses in one class, bringing down the 2051 ‘latent classes’ to a comprehensible, clearly distinguishable number of latent classes.

We used SAS Enterprise Guide 8.3 with the PROC LCA for the LCA analysis, considering the nested data structure. As the students were not required to answer all questions, 49 students of the 2051 (2.3%) left some tasks unanswered. Hot-deck imputation was used to impute these missing values.

Results

Overall response rates to the SMART-test *Meaning of Letters*

The overall response rates of the participants are shown in Figure 2. Note the low number of correct answers as well as that LU answers are only possible in task 1 and 2, while SAC answers are only possible in tasks 3 to 6. The LO misconception was omnipresent in the responses to the meaning and proportional tasks; while SAC was present in most responses to the additive tasks.

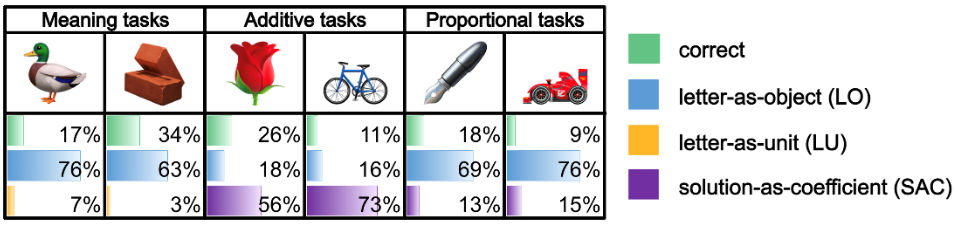


Figure 2: Overall response rates to the *Meaning of Letters* test

Model selection and model fit of LCA

We iteratively built several models with different numbers of classes to choose the appropriate number of classes for our LCA model. Information criteria (AIC/BIC) and the possibility of giving meaningful labels to the classes were used to decide the number of classes. The 6-class model had the best information criteria, the most interpretable latent classes, and no particularly infrequent class. Table 1 presents the given labels to each class and the prevalence (indicating overall class membership probability) of each class. There is no order or hierarchy of the classes.

			Mean classification probability					
Classes with labels		**Prevalence**	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Class 5**	**Class 6**
Class 1	LO predominant	23%	**85%**	2%	5%	1%	6%	1%
Class 2	LO with SAC	11%	4%	**91%**	1%	3%	2%	0%
Class 3	LO with SAC only in additive tasks	38%	6%	1%	**91%**	2%	1%	0%
Class 4	Correct meaning tasks with LO/SAC elsewhere	8%	4%	3%	9%	**77%**	3%	3%
Class 5	LO apart from additive tasks	15%	8%	0%	15%	2%	**70%**	4%
Class 6	Mostly correct	5%	3%	1%	1%	6%	8%	**81%**

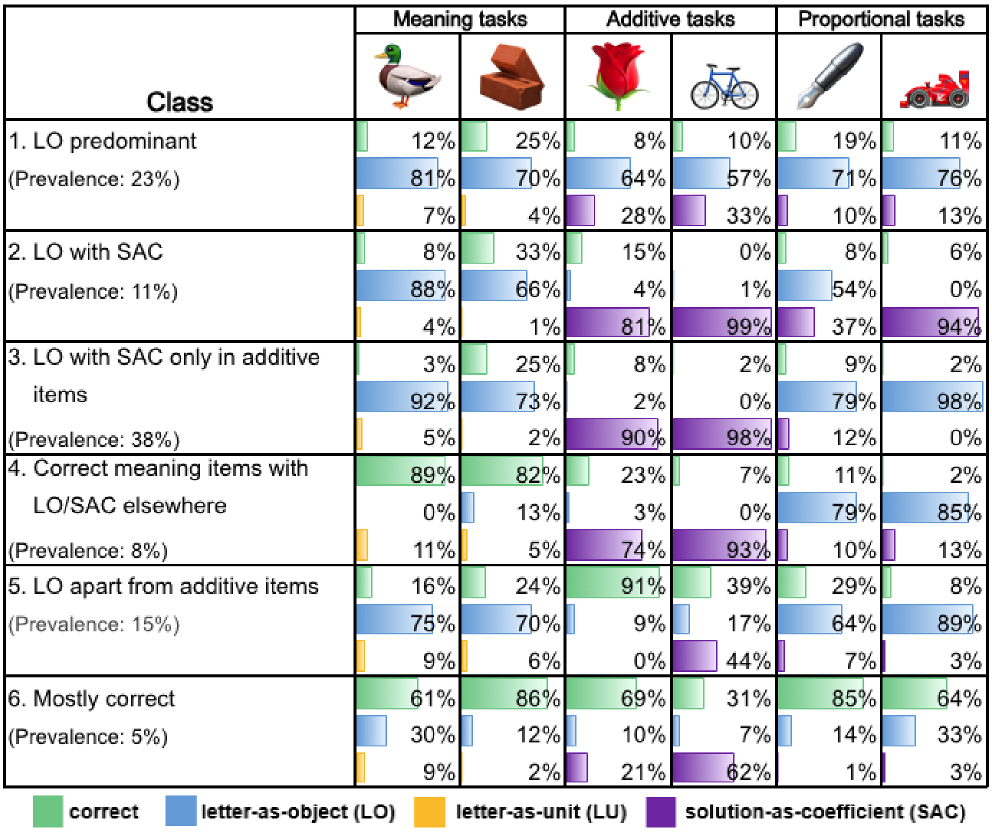
Table 1: Mean classification probability; hit rate in the diagonal (bold)

LCA allows the calculation of the likelihood that a student belongs to each class by analysing their test responses. A robust LCA model, characterised by high homogeneity among latent classes and clear class separation, yields for most students a class where the probability of classification is high for that best-fitting class and low for the others. By calculating the mean classification probability for all students aggregated along their best-fitting class, we get a grip on the homogeneity and class separation in our 6-class LCA model. These mean classification probabilities are shown in Table 1: for example, a student with class 2 as best has a probability of 91% to be in this class (we call this the 'hit rate', shown in the diagonal of the table) and only a probability of 4% to be in the first.

Description of the latent classes

In Figure 3, the item-response probabilities for each task are shown for the six classes. In the following paragraphs, we give a description of each of the classes.

Class 1 (prevalence: 23%) is characterised by LO responses being most likely in all tasks. Even when the initial letter of the involved object is not used (*Bricks*), LO is more likely than a correct answer. The subtype SAC is also possible in additive tasks, but less likely than LO, indicating a relatively consistent interpretation of letters as abbreviations for objects. We label this class "*LO predominant*"



Class	Meaning tasks		Additive tasks		Proportional tasks	
	🦆	🧱	🌹	🚲	🖋	🏎
1. LO predominant (Prevalence: 23%)	12% 81% 7%	25% 70% 4%	8% 64% 28%	10% 57% 33%	19% 71% 10%	11% 76% 13%
2. LO with SAC (Prevalence: 11%)	8% 88% 4%	33% 66% 1%	15% 4% 81%	0% 1% 99%	8% 54% 37%	6% 0% 94%
3. LO with SAC only in additive items (Prevalence: 38%)	3% 92% 5%	25% 73% 2%	8% 2% 90%	2% 0% 98%	9% 79% 12%	2% 98% 0%
4. Correct meaning items with LO/SAC elsewhere (Prevalence: 8%)	89% 0% 11%	82% 13% 5%	23% 3% 74%	7% 0% 93%	11% 79% 10%	2% 85% 13%
5. LO apart from additive items (Prevalence: 15%)	16% 75% 9%	24% 70% 6%	91% 9% 0%	39% 17% 44%	29% 64% 7%	8% 89% 3%
6. Mostly correct (Prevalence: 5%)	61% 30% 9%	86% 12% 2%	69% 10% 21%	31% 7% 62%	85% 14% 1%	64% 33% 3%

🟩 correct 🟦 letter-as-object (LO) 🟨 letter-as-unit (LU) 🟪 solution-as-coefficient (SAC)

Figure 3: The six latent classes with their item-response probabilities for every task

Class 2 (prevalence: 11%) is characterised by high probabilities for SAC responses. However, for *Biros* a LO response is more likely than a SAC response. In the meaning tasks that do not offer a SAC option, LO is most likely; a correct response in *Bricks* is possible. Since this is the class with the highest probability for SAC in proportional tasks, students in this class seem to be quite convinced that coefficients stand for (possible) solutions also in different equation types. We label this class "*LO with SAC*".

Class 3 (prevalence: 38%) is characterised by very high probabilities for SAC responses in additive tasks and high probabilities for LO responses in all other tasks. This indicates a rather consistent interpretation of letters as abbreviations for involved objects in combination with an interpretation of coefficients as solutions in additive equations. We label this class *"LO with SAC only in additive tasks*".

Class 4 (prevalence: 8%) is characterised by high probabilities for correct responses in meaning tasks, high probabilities for SAC in additive tasks, and high probabilities for LO in proportional tasks. Students in this class seem to be able to identify the correct meaning of an algebraic letter when directly being asked for it. However, when choosing equations they still fall into the trap of interpreting letters as abbreviations (and coefficients as solutions in additive equations). We label this class "*Correct meaning tasks with LO/SAC elsewhere*".

Class 5 (prevalence: 15%) is characterised by medium to high probabilities for LO in all tasks other than additive tasks. While in *Garden* the correct response is most likely, in *Wheels* SAC is slightly more likely than the correct response. For this response pattern, it is impossible to identify one reason (see Discussion). We label this class "*LO apart from additive tasks*".

Class 6 (prevalence: 5%) is characterised by correct responses being most likely in almost all tasks. Only in *Wheels*, SAC is double as likely as the correct response. In *Ducks* and *Racetrack*, LO is also possible but half as likely as the correct response. This indicates at least a partial understanding that algebraic letters do not stand for abbreviations. We call this class "*Mostly correct*".

Discussion and Outlook

Utilising LCA, six distinct response pattern classes were identified, offering detailed insights into the relationship between students' comprehension, misconceptions, and test tasks. These classes play a crucial role in enhancing our understanding of how students interpret algebraic letters across various contexts. It is important to note that a comprehensive understanding of the implications is an ongoing research process, and this discussion marks our initial attempt at exploring these insights.

Starting with classes that exhibit at least some correct answers, a notable discovery is that Class 6, characterised by mostly correct answers, has a low prevalence of 5% and still shows many SAC responses to the *Wheels* task. The absence of a class labelled 'All answers correct' is not surprising, as only 21 students (1%) would belong to this class, which contradicts the principle of a good LCA model that avoids very rare classes. Class 4 (8%) is intriguing, displaying high probabilities for correct meaning tasks, but struggles when translating this understanding into equations. These students seem to possess a superficial knowledge of variable meanings, adequate for direct inquiries about meaning with one variable but insufficient when dealing with equations involving two variables. This underscores the importance of recognising that merely asking about the meaning of letters in simple contexts does not necessarily imply a deep and accurate understanding. Class 5 (15%) is characterised by a very high probability of a correct answer on the *Garden* task and LO/SAC in most other tasks. Since these students do not seem to grasp the meaning of letters, it is likely that these correct responses are not a result of (partial) understanding but of a strategy of combining given letters and numbers according to the described situation without proper understanding.

Regarding the LO misconception and its subtypes, it is crucial to highlight that the LU misconception had a minimal occurrence in the meaning tasks. Some students consistently show LO (Class 1, 23%); however, even in this class, the subtype SAC has a probability of 33% in *Wheels*. This might indicate that this task especially fosters students' urge to come up with a numerical solution. In general, the subtype SAC is often clearly present in additive tasks only (especially Classes 3 and 4). Such additive equations can probably be read more intuitively as a solution sentence such as "35 bikes plus 10 trikes have 100 wheels altogether" compared to proportional tasks that would have to be read as something like "Sam bought 30 biros in 10 packs". It is also possible that the additive tasks more easily trigger some students' desire to give a solution than proportional tasks. In this respect, Class 2 is exceptional: the probability for SAC is very high in *Racetrack,* but only 37% in *Biros*, while they are both proportional tasks (with similar response rates, see Figure 2). This might indicate that a SAC interpretation in proportional equations is more likely when the letters involved refer to non-physical objects (e.g., rounds in *Racetrack)* or can be interpreted as units (e.g., minutes in *Racetrack).*

The analysis of the identified classes underlines how important the task type, complexity, and context – including the realness of involved entities and the underlying structure of the equation – seems to be for correctly interpreting algebraic letters. These are aspects that need to be taken into account for teaching as well as assessment. For example, problems that focus on a numerical interpretation (like "Think of a number") or require operations on both sides of the equation might help to support a correct interpretation of the equal sign as well as of algebraic letters. Two further questions remain for future research: How do students transition from these classes after a lesson series about algebraic letters? And: How can this analysis improve the SMART-test *Meaning of Letters* diagnosis?

Acknowledgements

This research was supported by a Travel Award from the Interdisciplinary Centre for Educational Research (IZfB) of the University of Duisburg-Essen granted to Katrin Klingbeil.

References

Akhtar, Z., & Steinle, V. (2017). The prevalence of the 'letter as object' misconception in junior secondary students. In A. Downton, S. Livy & J. Hall (Eds.), *Proceedings of the 40th annual conference of the Mathematics Education Research Group of Australasia* (pp. 77–84). MERGA.

Arcavi, A., Drijvers, P., & Stacey, K. (2017). *The learning and teaching of algebra: Ideas, insights, and activities*. Routledge.

Brandenburger, M., & Schwichow, M. (2023). Utilizing Latent Class Analysis (LCA) to Analyze Response Patterns in Categorical Data. In X. Liu, & W. J. Boone (Eds.), *Advances in Applications of Rasch Measurement in Science Education. Contemporary Trends and Issues in Science Education*, vol 57. (pp. 123 –154). Springer, Cham. https://doi.org/10.1007/978-3-031-28776-3_6

Clement, J., Lockhead, J., & Monk, G. (1981). Translation difficulties in learning mathematics. *American Mathematical Monthly*, *88*(4), 286–290.

Klingbeil, K., Rösken, F., Barzel, B., Schacht, F., Stacey, K., Steinle, V., & Thurm, D. (2024). Validity of multiple-choice digital formative assessment for assessing students' (mis)conceptions: Evidence from a mixed-methods study in algebra. *ZDM – Mathematics Education*. https://doi.org/10.1007/s11858-024-01556-0

Küchemann, D. (1981). Algebra. In K. M. Hart, M. L. Brown, D. E. Küchemann, D. Kerslake, G. Ruddock, & M. McCartney (Eds.), *Children's understanding of mathematics: 11-16* (pp. 102–119). John Murray.

MacGregor, M., & Stacey, K. (1997) Students' understanding of algebraic notation: 11-15. *Educational Studies in Mathematics*, *33*(1), 1–19.

Steinle, V., Gvozdenko, E., Price, B., Stacey, K., & Pierce, R. (2009). Investigating students' numerical misconceptions in algebra. In R. Hunter, B. Bicknell & T. Burgess (Eds.), *Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 491–498). MERGA.

The potential of AI: generating answers for multiple-choice questions using ChatGPT

Laura Kuusemets[1]

[1]University of Tartu, Faculty of Science and Technology, Tartu, Estonia; laura.kuusemets.1@ut.ee

Computer-based multiple-choice tests reduce teachers' workloads and enable students to receive immediate feedback on their performance, which is why they are widely used in the education system. To create a high-quality multiple-choice test, the author has to develop plausible distractors to be included among the provided answers. Therefore, creating a high-quality test is time-consuming. The purpose of this paper is to determine the suitability of ChatGPT for generating answer options for arithmetic and textual multiple-choice questions using Estonian and English prompts as examples, based on typical mistakes made by students and their similarities and differences with the answer options created by a human expert. This paper uses ChatGPT3.5 to test if it is possible to generate answers for multiple-choice questions that are based on given parameters. The results show that it is possible to generate answer options for multiple-choice questions using ChatGPT.

Keywords: Multiple-choice questions, artificial intelligence, mathematics education, distractors.

Introduction

The capacity of computers to execute cognitive tasks commonly associated with human intelligence, especially in learning and problem-solving, is defined as artificial intelligence (AI). AI-based systems are used to support teachers, reduce their workload, and automate assessment (Baker & Smith, 2019). In addition, AI-based assessment systems make the assessment process easier and faster for teachers (Kersting et al., 2014). Chat Generative Pre-Trained Transformer (ChatGPT) is an AI technology that generates conversational interactions based on user prompts (OpenAI et al., 2023). Large language models (LLM), such as ChatGPT have been pre-trained on huge volumes of textual data and are therefore able to answer questions with high accuracy, generate text and perform other language-related tasks (Kasneci et al., 2023). The ChatGPT model has proven its potential in various domains, including education (Liu et al., 2023).

Automated assessment and multiple-choice questions (MCQs) tools have been used for a long time. Because a trained language model can answer questions, it can be used to generate multiple-choice answers for tests. However, experience shows that the generation of a high-quality MCQ depends on the quality of the prompts (Kıyak, 2023). While it should not be overlooked, a study conducted by Liu et al., (2023) showed that ChatGPT has rather limited mathematical ability.

In the case of MCQs, the content of the question should be as concise, clear, precise, and unambiguous as possible, (Kelly, 1916); it should refer to the substance of the problem and the statement followed by the question to be answered. One answer option is always correct – it is called ’the key’ – and one or more answer options are always false, known as ‘decoy responses’ or ‘distractors’ (Gierl et al., 2017; McNichols et al., 2023). The answer options must not be partially true/false (Kelly, 1916). Distractors must be plausible and linked to the mistakes made by students and they should be misleading for students but not entirely false, which would make them easy to eliminate (Gierl et al., 2017; McNichols et al., 2023). If many distractors need be generated, this becomes burdensome for the test writer (Gierl et al., 2017).

Method

The main objective in using computer-based environments and MCQs, is to reduce teachers' workload through automated testing. Automated tests with multiple-choice answers are not widespread in the Estonian education practice, as preparing tasks and multiple-choice answers is time-consuming. Due to the aging teaching staff in Estonia, mathematics teachers are reluctant to use computer-based programs because they lack of digital competencies, and their command of English is relatively weak. If ChatGPT could generate answer options for MCQs in Estonian, the teacher would only have to enter the questions and answers for the test. As a result, teachers' workload would be significantly reduced, and teachers would have more time to support the students who need additional tutoring.

The purpose of this paper is to determine the suitability of ChatGPT for generating answer options for arithmetic and textual MCQs using Estonian and English prompts as examples, based on typical mistakes made by students and their similarities and differences with the answer options created by s human expert. In this study, both arithmetical and textual tasks are discussed, both arithmetic skills and knowledge of rules and theorems are checked in Estonian school mathematics. The topics and questions for the study have been taken from the textbook "Testid koolimatemaatikas I" (“Tests in school mathematics I”) by Lea Lepmann, a mathematics didactician at the University of Tartu. Following the textbook, we compare arithmetic and textual answer options generated by ChatGPT with those created by a human expert, a mathematics didactician, in terms of their accuracy and the quality of distractors and keys. When creating the prompt for ChatGPT, the recommendations for creating multiple-choice answers were taken into account, based on typical mistakes made by students. The ChatGPT prompt was entered in two languages – Estonian and English. To answer the research question, the answer options generated by ChatGPT and their correctness and the comparability of the false answers were analysed and compared with the answer options created by the mathematics didactician. ChatGPT3.5 has been used to carry out the study, and three research questions have been formuled:

1) How appropriate are ChatGPT-generated answer options, based on students' typical mistakes, for multiple-choice questions in an arithmetic task compared to human expert-generated answer options?

In order to an answer to the research question, four different types of arithmetic tasks were chosen for one subtopic of Lepmann’s (1991) textbook – addition fractions with same denominators, subtraction of fraction with mixed numbers, addition and subtraction of fractions with different denominator.

2) How appropriate are ChatGPT-generated answer options, based on students' typical mistakes, for multiple-choice questions in a textual task compared to human expert-generated answer options?

In order to answer the research question, two theory-based textual tasks were chosen from Lepmann's (1991) textbook – natural and prime numbers. The aim was to check whether ChatGPT understood the prompt and wether the output was formulated using correct mathematical vocabulary.

3) What are the differences in answer options between the English and Estonian prompts in ChatGPT?

ChatGPT prompt can also be in Estonian and can be used for understanding and constructing textual tasks in the local language, Estonian. To evaluate if ChatGPT can provide didactically correct answer options in the Estonian language in arithmetic and textual tasks, research questions 1 and 2 must be analysed.

Results

The results of each prompt are presented as a table, where each answer option is accompanied by its type: key (k), distractor (d), or incorrect option (i).

1) How appropriate are ChatGPT-generated answer options, based on students' typical mistakes, for multiple-choice questions in an arithmetic task compared to human expert-generated answer options?

Prompts 1.1-1.4: Prepare four answer options for the actions ________, where one answer option is correct, and three are incorrect. When preparing incorrect answer options, rely on typical mistakes made by students.

Prompt 1.1 – After receiving prompts in English and Estonian, ChatGPT prepared four answer options, one correct and three false (Table 1). There was a correct answer, in both instance and the generated distractor responses were plausible. The prompt given in Estonian, resulted in a non-reduced answer ($\frac{8}{34}$). For some of the answer options created by Lepmann, a partial calculation step has been added. There is also a non-reduced fraction ($\frac{8}{34}$).

Table 2. Four answers proposed by ChatGPT for action $\frac{5}{17}+\frac{3}{17}$

Input	Output		Lepmann, 1991
	In Estonian	In English	
$\frac{5}{17}+\frac{3}{17}$	$\frac{8}{17}$ (k) $\frac{15}{34}$ (d) $\frac{1}{10}$ (d) $\frac{8}{34}$ (d)	$\frac{8}{17}$ (k) $\frac{15}{17}$ (d) $\frac{2}{17}$ (d) $\frac{35}{17}$ (d)	$\frac{8}{17\cdot17}=\frac{8}{289}$ (d) $\frac{5\cdot17+3\cdot17}{17}=\frac{136}{17}$ (d) $\frac{8}{17}$ (k) $\frac{8}{34}$ (d)

Prompt 1.2 – After receiving a prompt in Estonian, ChatGPT generated four answer options, one correct and three false (Table 2), i.e., they match the given parameters. Two of the generated distractor answers were negative, and one option is in a non-reduced form. ChatGPT generated four answer options for the English prompt, all of which were false. One incorrect answer option was marked as the key. The wrong answer options were all in a non-reduced form. If the same prompt was given a second time, ChatGPT created new answer options and indicated the correct answer, but it was in a non-reduced form ($1\frac{2}{6}$). One option was also in a non-reduced form ($2\frac{2}{6}$), two other answer options

were not simplified correctly and were not reduced ($1\frac{8}{6}$ and $2\frac{8}{6}$). The answer options created by Lepmann show a partial calculation step. One of the answer options is in a non-reduced form ($2\frac{4}{6}$).

Table 3. Four answers proposed by ChatGPT for action $4\frac{1}{6} - 2\frac{5}{6}$

Input	**Output**			**Lepmann, 1991**
	In Estonian	**In English**		
$4\frac{1}{6} - 2\frac{5}{6}$	$1\frac{1}{3}$ (k) $-1\frac{1}{3}$ (i) $-\frac{4}{3}$ (i) $3\frac{2}{3}$ (d)	$1\frac{2}{3}$ (k) $1\frac{4}{6}$ (d) $1\frac{4}{36}$ (d) $3\frac{3}{12}$ (d)	$1\frac{2}{6}$ (k) $2\frac{2}{6}$ (d) $1\frac{8}{6}$ (d) $2\frac{8}{6}$ (d)	$(4-2)+\left(\frac{1}{6}-\frac{5}{6}\right)=2-\frac{4}{6}=1\frac{1}{3}$ (k) $(4-2)+\left(\frac{5}{6}-\frac{1}{6}\right)=2+\frac{4}{6}=2\frac{2}{3}$ (d) $(4-2)-\left(\frac{1}{6}-\frac{5}{6}\right)=2-\frac{1-5}{6}=2\frac{4}{6}$ (d) $(4-2)-\left(\frac{1}{6}+\frac{5}{6}\right)=2-\frac{6}{6}=1$ (d)

Prompt 1.3 – In response to a prompt in Estonian, ChatGPT generated three answer options, one correct and three false, i.e., matching the given parameters (Table 3). If the prompt was given in English, ChatGPT did not produce any answer options that matched the parameters. A wrong answer was marked as the key, and there was no correct answer option. When the same prompt was given a second time, the results were partially different; again, a wrong answer was marked as the key, and a correct answer option was missing. In both cases, there was no correct answer among the outputs. If ChatGPT was prompted to solve an action, it returned the correct answer with solution steps. Only then, it generated answer options corresponding to the parameters, and the correct answer option was in a reduced form. Three of the four answer options created by Lepmann show a partial calculation step. Only one out of the four answer options was in a non-reduced form ($\frac{7}{21}$).

Table 4. Four answers proposed by ChatGPT for action $\frac{2}{3} + \frac{5}{7}$

Input	**Output**				**Lepmann, 1991**
	In Estonian	**In English**			
$\frac{2}{3} + \frac{5}{7}$	$\frac{29}{21}$ (k) $\frac{7}{10}$ (d) $\frac{10}{21}$ (d) $\frac{34}{21}$ (d)	$\frac{31}{21}$ (d) $\frac{1}{7}$ (d) $\frac{10}{21}$ (k) $\frac{7}{10}$ (d)	$\frac{31}{21}$ (k) $\frac{7}{10}$ (d) $\frac{15}{21}$ (d) $\frac{2}{10}$ (d)	$1\frac{8}{21}$ (k) $\frac{7}{10}$ (d) $\frac{29}{21}$ (d) $\frac{5}{7}$ (d)	$\frac{2\cdot7+3\cdot5}{3+7}=\frac{29}{10}$ (d) $\frac{2\cdot7+3\cdot5}{21}=\frac{29}{21}$ (k) $\frac{2+5}{21}=\frac{7}{21}$ (d) $\frac{7}{10}$ (d)

Prompt 1.4 – Having receiveda prompt in Estonian, ChatGPT suggested four answer options, one correct and three false (Table 4). The output matched the parameters. The distractors are all in a reduced form. When providing a prompt in English, the output was incorrect the first time. The output was similar to the output given in Estonian and to the answer options produced by Lepmann. However, a wrong answer option was marked as correct and it was in a non-reduced form. In response to a prompt requiring ChatGPT to solve the given operation first and then generate the answer options, it produced the correct answer option in a reduced form. Two of the generated false answers were

equal – one in a reduced and the other in a non-reduced form ($\frac{2}{6}$ and $\frac{1}{3}$). Another distractor was also in a non-reduce form ($\frac{3}{12}$). All of the answer options created by Lepmann include calculation steps.

Table 5. Four answers proposed by ChatGPT for action $\frac{2}{3}-\frac{1}{6}$

Input	Output			Lepmann, 1991
	In Estonian	In English		
$\frac{2}{3}-\frac{1}{6}$	$\frac{1}{2}$ (k)	$\frac{1}{6}$ (d)	$\frac{1}{2}$ (k)	$\frac{2-1}{6}=\frac{1}{6}$ (d)
	$\frac{1}{3}$ (d)	$\frac{1}{3}$ (k)	$\frac{2}{6}$ (d)	$\frac{4-1}{3}=\frac{3}{3}=1$ (d)
	$\frac{1}{6}$ (d)	$\frac{3}{6}$ (d)	$\frac{3}{12}$ (d)	$\frac{4-1}{6}=\frac{3}{6}=\frac{1}{2}$ (k)
	$\frac{3}{4}$ (d)	$\frac{2}{9}$ (d)	$\frac{1}{3}$ (d)	$\frac{12-3}{6}=\frac{9}{6}=\frac{3}{2}$ (d)

2) How appropriate are ChatGPT-generated answer options, based on students' typical mistakes, for multiple-choice questions in a textual task compared to human expert-generated answer options?

Prompts 2.1-2.4: Prepare four answer options for the assertion "______", where one is correct and three are incorrect. When preparing incorrect answer options, rely on typical mistakes made by students.

Prompt 2.1 – After receiving a prompt in Estonian, ChatGPT produced an output partially corresponding to the parameters (Table 5). The marked correct answer is partially true because the set of natural numbers is closed under addition and multiplication but not under subtraction and division. For the English prompt, ChatGPT gave an output that partially matches the parameters, indicating addition as the correct answer while noting subtraction and division as incorrect.

Table 6. Options proposed for the assertion "The set of natural numbers is closed …"

Input	Output		Lepmann, 1991
	In Estonian	In English	
"Naturaalarvude hulk on kinnine ..." *„The set of natural numbers is closed ..."*	• Liitmise ja lahutamise suhtes (*Under addition and subtraction); (k)* • Jagamise suhtes (*Under division); (d)* • Murdude suhtes (*Under fractions); (d)* • Korrutamise suhtes (*Under multiplication). (d)*	• Under addition; (k) • Under subtraction; (d) • Under multiplication; (d) • Under division. (d)	• Ainult liitmise suhtes (*Only under addition); (d)* • Ainult liitmise, lahutamise ja korrutamise suhtes (*Only under addition, subtraction); (d)* • Ainult liitmise ja korrutamise suhtes (*Only under addition, multiplication); (k)* • Kõigi nelja aritmeetilise tehte suhtes (*Under all four arithmetical operation). (d)*

Prompt 2.2 – In response to a prompt in Estonian, ChatGPT generated an output corresponding to the parameters (Table 6). For English prompts, however, ChatGPT solved the task completely differently by generating true/false statements. Each answer option explains why the statement is true/false and the reasoning is relevant.

Table 7. Options proposed for the assertion “Of the following numbers, 1, 2, 3, 4, 5, 6, 7, 8, 9 are prime numbers"

Input	Output		Lepmann, 1991
	In Estonian	In English	
"Järgmistest arvudest 1, 2, 3, 4, 5, 6, 7, 8, 9 on algarvud" *"Of the following numbers 1, 2, 3, 4, 5, 6, 7, 8, 9 are prime numbers"*	• 2, 3, 5 ja 7; (k) • 1, 2, 3, 5, ja 7; (d) • 2, 4, 6, ja 8; (d) • 2, 3, 4, ja 7. (d)	• False. 1, 4, 6, 8 and 9 are not prime numbers (k) • True. All the numbers are prime; (d) • False. Only 2, 3, 5, 7 are prime numbers; (d) • True. Any single-digit number is a prime number. (d)	• Ainult 1, 3, 5, 7 (*only 1, 3,5, 7); (d)* • Ainult 2, 3, 5, 7 (*only 2, 3, 5, 7); (k)* • Ainult 1, 3, 5, 7, 9 (*only 1, 3, 5, 7, 9); (d)* • Kõik nimetatud arvud (*All given numbers). (d)*

Discussion

The study results showed that when constructing multiple answers to given operations, the wording of the parameters in the prompt requires careful consideration. ChatGPT generated multiple-choice answers based on the parameters.

From the answer options generated by ChatGPT it is evident that they still require human verification. The generated multiple-choice answers included many mistakes, such as marking a wrong answer option as the correct one, especially with the English prompt. With the Estonian prompt, ChatGPT generated the correct answer option and three false options for all the given tasks on the first run. Most of the false responses were also plausible, which means that ChatGPT can generate suitable answer options in response to an Estonian prompt. Comparing the answer options generated by ChatGPT with the options generated by the mathematic didactician Lea Lepmann, there are some similarities for some tasks but no complete overlaps. For some tasks, no lure responses were the same for the Estonian prompt, the English prompt, and the response options of the human expert (Table 2). However, for some tasks, there were response options that occurred in all three (Table 3, 4). With an English prompt, ChatGPT marked an incorrect answer option as the key 75% of the time on the first run. Most of the false responses were also plausible, which means that in the case of an English prompt, the answer options generated by ChatGPT may not be suitable and need to be checked by teacher. ChatGPT provides suitable answer options if it first solves the action. An important difference is that the answer options generated by the human expert are structurally different – they include calculation steps that represent the mistake made in the calculation of fraction. None of the answer options generated by ChatGPT included calculation steps.

Care must be taken to ensure that the answer options and the task statement are consistent with each other and the curriculum. In the topic of common fractions, the problem statement should indicate the expected form of the answer (reduced or non-reduced). Estonian school mathematics requires the answer to be given in a reduced form. In this case, this should also be described in the stem. In Estonian school mathematics, fractions are taught in the 6th grade, while calculation with negative numbers is taught in the 7th grade. Therefore, two answer options can be eliminated from the Estonian output for Prompt 1.2. If an answer option contains an answer in a form that the students have not yet learned, they are likely to treat this option as incorrect. According to Girel et al., (2017) an answer option must not be completely wrong because in this case, students will be able to eliminate the answer immediately.

In Prompt 2.1, one can see that ChatGPT makes mathematical mistakes and seems that ChatGPT is aware of the mistaks but still marks the wrong answer as the key. The testing conducted in this study showed that ChatGPT can understand the content of the text and generate a textual response in both Estonian and English. The answers are plausible in Estonian but require linguistic editing. When comparing the answer options in Estonian and English and the one prepared by the didactician, they are similar, or only one of the answer options is different. The most significant difference was for Prompt 2.2, where ChatGPT produced true/false statements for the English prompt, which is undesirable for MCQs because it intentionally complicates the test.

Nevertheless, there were errors or omissions in compiling the multiple-choice answers. For several of the statements, ChatGPT gave a wrong or incomplete answer. For both English and Estonian prompts, there were errors in the mathematical terms or rules in the output. When generating answer options for arithmetical tasks, ChatGPT performed better with Estonian prompts than with English prompts. There were many errors in the outputs in response to English prompts, and the correct answer option often needed to be corrected. Thus, ChatGPT is good at generating multiple-choice answers based on Estonian prompts. Furthermore, it was found that ChatGPT can understand Estonian prompts and produce verbal responses. Based on the results, Estonian teachers can use ChatGPT to create multiple-choice answers, but it does not necessarily result in efficiency gain for teachers, as all the multiple-choice answers still have to be checked by the teacher.

Conclusion and future work

In this work, ChatGPT was used to see if it could make it easier and faster for teachers to create multiple-choice answers. The most important findings are that, first, the variability of the distractors generated by ChatGPT is smaller than the variability of the distractors generated by mathematical didactician. Second, the keys generated by ChatGPT are incorrect in some cases incorrect. In the case of Estonian prompts, ChatGPT failed to give a correct answer only in a textual task where it marked a partially false statement as the key. With English prompts for arithmetic tasks, ChatGPT marked the correct answer option as the key only 25% of the time in the first attempt. In the case of textual tasks, it indicated the wrong answer option as the correct one in some of the outputs. The third and most important result for Estonian mathematic teachers was that ChatGPT generates more evenly matched responses for Estonian prompts than for English prompts. In some cases, the distractors generated in Estonian are more plausible than those generated in English, because there were more answer-options that were non-reduced fractions.

The present work's limitations lie in using an older version of ChatGPT, 3.5, when the newer version 4.0 is more advanced, has newer data, and is better at processing textual information. In this work, version 3.5 is used because it is available free of charge to everyone, while 4.0 is a paid version. Even though ChatGPT 4.0 is not free of charge, would be advisable to use the newer version.

In future work, ChatGPT should be given more details on the desired learning outcomes for a given topic are on students' common misconceptions in that topic. It should then be investigated whether ChatGPT can generate more appropriate distractor responses based on the given parameters and, for each false response, generate feedback that supports the student based on the type of mistake.

References

Baker, T., & Smith, L. (2019). *Educ-AI-tion Rebooted? Exploring the future of artificial intelligence in schools and colleges*. https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.pdf

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research*, *87*(6), 1082–1116. https://doi.org/10.3102/0034654317726529

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kelly, F. J. (1916). The Kansas Silent Reading Tests. *Journal of Educational Psychology*, *7*(2), 63–80. https://doi.org/10.1037/h0073542

Kersting, N. B., Sherin, B. L., & Stigler, J. W. (2014). Automated Scoring of Teachers' Open-Ended Responses to Video Prompts: Bringing the Classroom-Video-Analysis Assessment to Scale. *Educational and Psychological Measurement*, *74*(6), 950–974. https://doi.org/10.1177/0013164414521634

Kıyak, Y. S. (2023). A ChatGPT Prompt for Writing Case-Based Multiple-Choice Questions. *Revista Española de Educación Médica*, *4*(3). https://doi.org/10.6018/edumed.587451

Lepmann, L. (1991). *Testid koolimatemaatikast I*. Eesti õppekirjanduse keskus.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology*, *1*(2), 100017. https://doi.org/10.1016/j.metrad.2023.100017

McNichols, H., Feng, W., Lee, J., Scarlatos, A., Smith, D., Woodhead, S., & Lan, A. (2023). *Exploring Automated Distractor and Feedback Generation for Math Multiple-choice Questions via In-context Learning* (arXiv:2308.03234). arXiv. http://arxiv.org/abs/2308.03234

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). *GPT-4 Technical Report*. https://doi.org/10.48550/ARXIV.2303.08774

Secondary mathematics teachers' experiences of using ChatGPT to design probability and statistics assessment items

Minsung Kwon[1] and Inah Ko[2]

[1]California State University Northridge, USA; minsung.kwon@csun.edu

[2]University of Michigan, USA; inahko@umich.edu

The purpose of this study was to explore secondary mathematics teachers' experiences of using ChatGPT to design probability and statistics assessment items. For this purpose, we analyzed 22 secondary mathematics teachers' conversations with ChatGPT and their survey responses in terms of their overall experiences with ChatGPT, their intentions to use ChatGPT-generated assessment items, and affordances and challenges of using ChatGPT to design assessment items. The results showed that most teachers did not specify the purpose of assessment and only one teacher identified mathematical errors in the ChatGPT's responses. The teachers employed a wide range of follow-up questions in responding to the ChatGPT's suggestions. The survey results showed that their intentions to use ChatGPT were polarized. The teachers perceived that ChatGPT provided affordances such as creativity and efficiency but shared their concerns about mathematical errors, inaccuracy, ethical issues, and its security.

Keywords: Summative assessment, probability and statistics, Artificial Intelligence (AI), ChatGPT, secondary mathematics teachers

Introduction

Since Shulman's seminal work on the pedagogical content knowledge (PCK) as a special amalgam of content and pedagogy needed for teaching (Shulman, 1986), the scholarship in teacher education has made a major shift from identifying teacher characteristics toward conceptualizing PCK, developing instruments to measure PCK, and designing teacher education programs to develop PCK. Besides the efforts made in the subject-matter specific PCK, researchers have identified the knowledge needed for teaching with technology in response to the emergence of new digital technology and the importance of technology competence. This specialized knowledge that teachers need to teach with technology is conceptualized as Technological, Pedagogical, and Content Knowledge (TPACK) and it is further specified into sub-domains (Misha & Koehler, 2006). After the publication of the *Second Handbook of Technological Pedagogical Content Knowledge (TPACK) for Educators* in 2016, we have experienced a rapid change and social demands of incorporating technology in teaching and learning over the past few years. Especially, the emergence of generative Artificial Intelligence (AI) and the public release of ChatGPT on November 30, 2022 have attracted more attention from educators, both for the benefits of using AI and concerns about using AI in teaching and learning.

Given these rapid changes using technology, we aimed to explore the experiences of secondary mathematics teachers to design probability and statistics assessment items using ChatGPT in Korea. Because of the nature of grading on a curve and its importance for the college admission in Korea, many secondary mathematics teachers have experienced challenges in writing assessment items that can result in a desired item difficulty and item discrimination but can be differentiated from commercially available workbooks. The advancement of ChatGPT made us to wonder how secondary mathematics teachers might use ChatGPT to design assessment items, what experiences they have with ChatGPT, and whether ChatGPT might resolve their persistent challenges or create new issues to write assessment items. More specifically, this paper examines the following research questions:

1. What prompts do secondary mathematics teachers use to design probability and statistics assessment items in ChatGPT?
2. What probability and statistics assessment items do ChatGPT generate? To what extent do teachers have intentions to use these ChatGPT-generated assessment items?
3. What affordances and challenges do secondary mathematics teachers perceive in using ChatGPT to design probability and statistics assessment items?

Methods

Using a convenience sampling method (Pattern, 1990), we collected the data from 22 secondary mathematics teachers who enrolled in a three-credit graduate course in mathematics education in Korea. After the instructor's short introduction of ChatGPT, the teachers were instructed to have conversations with ChatGPT 3.5 approximately for 10-15 minutes to design probability and statistics assessment items for Grade 11. We asked teachers to design an assessment item for probability and statistics, a topic that tends to require less use of advanced mathematical expressions or graphical representations. After conversations with ChatGPT, teachers were asked to complete a short survey including a URL link for their ChatGPT conversations, one final ChatGPT-generated probability and statistics assessment item, their overall experiences with ChatGPT, and whether ChatGPT understood their questions, intentions, and feedback. Additionally, we asked whether they had any mathematical or pedagogical issues with ChatGPT, the benefits and challenges of using ChatGPT for assessment, and whether they had intentions to use ChatGPT-generated assessment items using a five-point Likert scale (1: strongly disagree and 5: strongly agree). Lastly, the teachers were asked to explain their perceived item difficulty generated by ChatGPT and to evaluate its appropriateness for their students.

First, we analyzed the prompts that teachers used in ChatGPT to design probability and statistics assessment items. We asked the teachers to share a URL link for their ChatGPT conversations. At the time of analyzing the data, three links were invalid which resulted in 19 links for the analysis. In analyzing the teachers' ChatGPT conversations, it became clear that the strategy and skill of providing effective prompts are crucial for obtaining desirable responses. This is because ChatGPT generates responses based on the user's inputs or prompts. Some reports suggest that ChatGPT has often provided unvalidated and incorrect information (Einarsson et al., 2023). Considering that ChatGPT's mathematical skills are not as strong as in other areas (Frieder et al., 2023), it is important for the users, in this case secondary mathematics teachers, to employ the strategy of using effective prompts to get the desirable outcomes and assess the validity, accuracy, and credibility of ChatGPT's responses. In this paper, we categorized the teachers' prompts into four categories: 1) Specificity of the prompts to design an assessment item; 2) Identification of errors in ChatGPT's responses; 3) Quality of follow-up questions; 4) Evaluation of ChatGPT's responses. Table 1 illustrates our coding rubric to analyze teachers' prompts to ChatGPT.

Table 1. Coding rubric for teachers' prompts to ChatGPT in designing assessment items

Category	Description	Scoring Rubric
1.Specificity of prompts	Does a teacher's prompt identify... ● a topic? ● a difficulty level? ● a target grade-level? ● a purpose?	● 1: Address one of the sub-categories ● 2: Address two of the sub-categories ● 3: Address three of the sub-categories ● 4: Address four of the sub-categories
2.Identifying an error in ChatGPT's responses	Does a teacher identify… ● a contextual error? ● a mathematical error?	● 0: neither identified contextual nor mathematical error ● 1: identified either contextual or mathematical error ● 2: identified both contextual and mathematical error
3.Quality of follow-up questions	● Is a teacher's follow-up question based on interpretation or evaluation of the ChatGPT's responses? ● Does a teacher's follow-up question further specify or challenge the ChatGPT-generated item?	● 0: accept ChatGPT's responses without posing any follow-up questions or asking unrelated questions to ChatGPT's responses ● 1: ask a follow-up question to clarify ChatGPT's responses or request simple modification (e.g., different format) ● 2: ask a follow-question to further specify ChatGPT's responses by adding conditions or related concepts ● 3: ask a follow-up question to challenge the ChatGPT's responses or to provide a specific feedback to ChatGPT
4.Evaluating ChatGPT's responses	● Does a teacher evaluate ChatGPT's responses?	● 0: no ● 1: yes

To analyze the teachers' intentions to use ChatGPT-generated assessment items, we recoded disagree responses (1:strongly disagree and 2:somewhat disagree) to negative and recoded agree responses (4:somewhat agree and 5:strongly agree) to positive. For the affordances and challenges of using ChatGPT to design assessment items, we repeatedly read the teachers' responses and found themes emerged from their open responses using an inductive coding.

Results

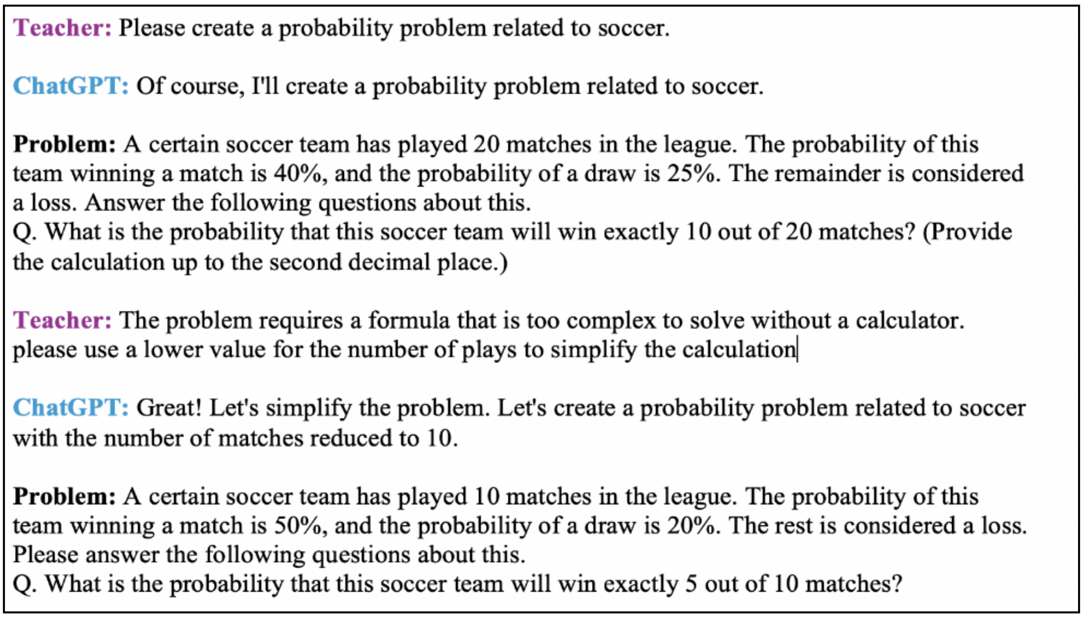
RQ1. Teachers' Prompts Used in ChatGPT

The first analysis focused on determining whether the prompts included key details such as a topic, difficulty level, target grade-level, and purpose (Category 1). The analysis revealed that 14 out of 19 teachers (74%) received a score of 2 for this category. This score indicates that they provided two components of their desired assessment items, such as a broad topic and target grade-level, without specifying the purpose for using the assessment item (e.g., "Please provide a problem about probability and statistics for high school students"). Only one of the 19 teachers provided all the four

components, including the goal of an assessment item ("Please rephrase the problem conditions to assess if a student can derive the solution by using a normalization.")

Second, the analysis of identifying a contextual or mathematical error (Category 2) suggested that only one out of 19 teachers explicitly mentioned an error in ChatGPT’s responses. For example, the teacher identified a mathematical error in ChatGPT's explanation of the suggested problem. Specifically, the teacher stated, "You mentioned that this problem will help students understand the concept of conditional probability, but your suggested problem is not related to the concept.”

Third, an analysis of the quality of follow-up questions (Category 3) indicated that five teachers (26%) either accepted the ChatGPT’s suggestions (e.g., “Good. Thank you!”) or asked unrelated questions to the ChatGPT’s responses (e.g., “Please give me another question”). Additionally, five teachers (26%) asked a simple parallel question (e.g., “Please give me a more difficult question”); five teachers (26%) further specified the ChatGPT’s responses (e.g., "Please add a given that uses the concept of conditional probability"); while four teachers (21%) either provided specific revisions (e.g., "Please use a lower value for the number of plays to simplify the calculation") or challenge the ChatGPT’s responses (e.g., "Isn't it possible for people with assigned seats to move among them?"). Figure 1 illustrates an example of one teacher’s ChatGPT conversation. As shown in Figure 1, the teacher provides specific feedback on ChatGPT-generated assessment item. In response to this feedback, ChatGPT revised the initial assessment item, using a smaller number for matches and different probability values. However, solving the problem would still require a calculator.



Teacher: Please create a probability problem related to soccer.

ChatGPT: Of course, I'll create a probability problem related to soccer.

Problem: A certain soccer team has played 20 matches in the league. The probability of this team winning a match is 40%, and the probability of a draw is 25%. The remainder is considered a loss. Answer the following questions about this.
Q. What is the probability that this soccer team will win exactly 10 out of 20 matches? (Provide the calculation up to the second decimal place.)

Teacher: The problem requires a formula that is too complex to solve without a calculator. please use a lower value for the number of plays to simplify the calculation

ChatGPT: Great! Let's simplify the problem. Let's create a probability problem related to soccer with the number of matches reduced to 10.

Problem: A certain soccer team has played 10 matches in the league. The probability of this team winning a match is 50%, and the probability of a draw is 20%. The rest is considered a loss. Please answer the following questions about this.
Q. What is the probability that this soccer team will win exactly 5 out of 10 matches?

Figure 1. An example of a teacher’s ChatGPT conversation to design a probability item

This teacher also received high scores in other categories. Specifically, the teacher clarified the purpose of an assessment item and evaluated ChatGPT-generated assessment item in terms of its applicability in real-life situations and item difficulty.

Finally, we investigated if there was clear evidence that teachers evaluated responses from ChatGPT. The analysis indicated that seven out of 19 teachers (37%) assessed the correctness and validity of ChatGPT-generated assessment item. For example, one of the prompts stated that "Your problem necessitates overly complex calculations to determine the standard deviation.”

RQ2. ChatGPT-generated assessment items and teachers' intentions to use them

For the second research question, we analyzed ChatGPT-generated assessment items that teachers came up with and their intentions to use these assessment items. As discussed in the previous section, teachers employed a wide range of prompts to design an assessment item which resulted in a wide range of assessment items. Table 1 illustrates four teachers' ChatGPT-generated probability and statistics assessment items with their intentions to use these items for the assessment (1: strongly disagree and 5: strongly agree).

Table 1. Examples of ChatGPT-generated assessment items and teachers' intention to use them

Assessment Item	Intention
Teacher A's ChatGPT-generated assessment item: A class of 10 students sits at a round table with 12 chairs. Two specific students, A and B, must sit next to each other, and two other students, C and D, must sit facing each other. The remaining six students can sit at random. Find the number of cases in which the students are seated at the round table.	4
Teacher B's ChatGPT-generated assessment item: Jimin and Junho are playing a coin flip game. The coin is assumed to be a fair coin, and the probability of getting heads (H) and tails (T) is 1/2 each. The rules of the game are as follows: 1. Jimin and Junho each flip a coin. 2. If it lands on heads (H), they win; if it lands on tails (T), they lose. 3. When both friends start flipping at the same time, Jimin flips first. Answer the following questions: 1. What is the probability that Jimin wins and what is the probability that Junho wins? 2. What is the probability that Jimin and Junho both win the first two coin tosses? 3. What is the probability that Jimin and Junho flip a total of 5 coins and win exactly 3 of them?	2
Teacher C's ChatGPT-generated assessment item: A soccer team has played 15 games in a league. The team has a 60% chance of winning a game and a 10% chance of drawing a game. The rest of the games are considered losses. Answer the following questions: 1. What is the probability that the soccer team will win exactly 10 out of 15 games? 2. What is the probability that this soccer team will win at least 12 out of 15 games? 3. What is the probability that this soccer team will lose or tie at least 5 out of 15 games?	4
Teacher D's ChatGPT-generated assessment item: You have a store that sells goods and you have five different types of goods. You need to display these items in a row, but the store's shelves are circular, so the first and last items are next to each other, i.e., they are arranged in a circle. Find the number of cases in which the store displays the five products in a circle.	2

Teacher A had an intention to use the ChatGPT-generated assessment item (rated 4 in a five-point Likert scale) but would like to revise the item by changing the number of chairs from 12 chairs to 10 chairs. However, Teacher A did not further explain the justification for this revision. Teacher B did not have an intention to use the ChatGPT-generated assessment item (rated 2 in a five-point Likert scale) because the ChatGPT-generated assessment item is simple, easy, and different from the expected item difficulty. Teacher B also pointed out that the assessment item has some inaccurate statements. Teacher C had an intention to use the ChatGPT-generated assessment item (rated 4 in a five-point Likert scale) but would like to revise the assessment item for the security reason. Teacher D did not have an intention to use the ChatGPT-generated assessment item (rated 2 in a five-point Likert scale) because it is quite similar to examples provided in textbooks, so it is more efficient for teachers to write their own assessment items.

In the survey, teachers have polarized responses about their intentions to use the ChatGPT-generated assessment items. Among 22 teachers, 10 teachers (45.5%) indicated that they had intentions to use the ChatGPT-generated assessment items, whereas the same number of teachers (45.5%) indicated that they did not have intentions to use the ChatGPT-generated assessment items. Two teachers (9%) responded neutrally about their intentions to use the ChatGPT-generated assessment items. Teachers who did not have intentions to use these assessment items explained that the ChatGPT-generated assessment items are too simple and easy, are not aligned well with their instruction, are not aligned with the curriculum, have inaccurate expressions, have incorrect answers, do not produce items with the intended item difficulties, are very similar to the textbooks, and are not quite different from items they can make. Two teachers wrote:

> I have very little intention of using it; the difficulty level is very low; the solutions it presents may not be what the curriculum intends; and I don't think ChatGPT is up to the task of developing items that accurately assess the competencies that the curriculum wants students to develop.
>
> I don't intend to use it yet, because the problems that ChatGPT creates are very similar to the textbooks. Even if I kept asking for new problems, they would just repeat the first problem with different numbers. In other words, I don't think I would use ChatGPT specifically because it only gives me typical problems from the textbooks or problem sets I have.

RQ3. Affordances and Challenges of using ChatGPT to design assessment items

In analyzing the teachers' survey responses, we found that the teachers perceived the affordances of using ChatGPT in terms of creativity (rich ideas, new types of problems, various contexts, extending teacher's limited thinking, and reducing the pains of creating items), efficiency (cost, time, and speed), specific difficult levels (easy or medium-level difficulty), specific type of assessment (performance assessment or formative assessment), specific type of items (multiple-choice items), providing solutions to the problems, diagnosis of errors, creating scoring rubrics, convenience, neutral (excluding teachers' own biases or preferences), producing anticipated solutions, and using ChatGPT for students' learning. Among these affordances, nine teachers (41%) identified creativity and five teachers (23%) identified efficiency in cost, time, and speed. However, the teachers identified its challenges as potential mathematical errors (e.g., incorrect answers, inaccurate solutions, and vague expressions), misalignment with curriculum, ethical issues (e.g., copyright issues, lack of information about sources), possibility of teacher's heavy reliance on ChatGPT, and security issues. In addition to these potential limitations, the teachers also identified that ChatGPT does not reflect their interactions with their students during their lessons, insufficient database (not quite different from commercially available workbooks or textbooks), teacher's intention for the assessment, the validity of assessment created by ChatGPT, inappropriate item difficulty, and insufficient item discrimination. Among these limitations, eight teachers (36.4%) identified the potential mathematical errors and seven teachers (31.8%) concerned about the security of assessment because of its nature of open source.

Discussions

The purpose of this study was to explore secondary mathematics teachers' experiences of using ChatGPT to design probability and statistics assessment items. For this purpose, we analyzed secondary mathematics teachers' conversations with ChatGPT and their survey responses about overall experiences with ChatGPT, their intentions to use ChatGPT-generated assessment items, and affordances and challenges of using ChatGPT in designing assessment items in Korea. The first research question examined the prompts that the secondary mathematics teachers used to design probability and statistics assessment items. The results show that most teachers specified some of the key details of the assessment items and did not identify mathematical errors or issues in ChatGPT's responses. On the other hands, the quality of follow-up questions is widely ranged (five teachers received score 0, five teachers received score 1, five teachers received score 2, and four teachers received score 3) and seven teachers evaluated ChatGPT's responses. The second research question examined the ChatGPT-generated probability and statistics assessment items after the teachers' conversations with ChatGPT and their intentions to use these assessment items. The survey results showed that the teachers had polarized responses to their intentions to use. The third research question examined what secondary mathematics teachers perceived its affordances and challenges of using ChatGPT in designing probability and statistics assessment items. Teachers perceived that ChatGPT provides affordances such as creativity and efficiency but identified limitations of potential mathematical errors, ethical issues, misalignment with curriculum, misalignment with instruction, and security of assessment because of its nature of open source.

The results of this study provide implications as follows. First, the analysis of teachers' prompts to design assessment items using ChatGPT provides implications that teachers need to be exposed to different types of prompts they can use in ChatGPT and they need to explore that the prompts they employed would determine whether they could get the desired outcomes. Without specific prompts, challenges, or evaluation of ChatGPTs' responses, some teachers simply accepted the ChatGPT's responses or repeated the same prompts to ChatGPT. Especially, many teachers mentioned that ChatGPT generated too easy items but did not produce the items with the intended item difficulty and sufficient item discrimination. As an exception, one teacher, who rated the ChatGPT-generated item as difficult, mentioned that ChatGPT was able to produce a more difficult item once the teacher added more conditions to the initial ChatGPT-generated item. In this study, we did not provide specific prompts that teachers can use in ChatGPT because we aimed to explore the quality of prompts that teachers use. However, we might offer examples of different types of prompts that teachers can use and then analyze the frequency of using specific prompts or explore how teachers employ different types of prompts to the same ChatGPT's responses.

Second, the teachers perceived that ChatGPT was creative, efficient, and convenient as affordances but addressed mathematical errors, inaccuracy, incorrectness, and vagueness as its major challenges. It is interesting to observe that many teachers identified these mathematical issues in the survey but few of them actually addressed these concerns in their conversations with ChatGPT. Facing such issues in ChatGPT, teachers should be able to use their mathematical knowledge to critically examine the mathematical accuracy, correctness, and performance of ChatGPT and address them to ChatGPT. Another challenges of using ChatGPT in designing assessment items are the misalignment between ChatGPT-generated items and curriculum in Korea and misalignment between ChatGPT-generated items and their own instruction. Because of its importance to use the specific grade-level mathematical vocabulary, concepts, or ideas outlined in the curriculum, it needs to be further examined whether ChatGPT understands or has an access to the specific curriculum materials and grade-level expectations in each country.

Lastly, writing assessment items is very stressful for teachers in Korea because students and parents are very sensitive to the assessment items and often complain if there are any vagueness, errors, or issues in the assessment and if there are any similarities between assessment items and commercially workbook items. Because the security of assessment items is one of the most important issues for teachers in Korea, they might not use the ChatGPT-generated assessment items without any major revisions or modifications. However, as three teachers commented in the survey, it would be a great learning opportunity for students to design or discuss a mathematical problem using ChatGPT. Because teachers experienced that ChatGPT often produced inaccurate, incorrect, and vague mathematical ideas, they would like to use ChatGPT-generated items as formative assessment for their students to explore whether students can detect any mathematical errors or issues in ChatGPT's responses.

References

Einarsson, H., Lund, S. H., & Jónsdóttir, A. H. (2023). Application of ChatGPT for automated problem reframing across academic domains. *Computers and Education: Artificial Intelligence*, 100194.

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., & Berner, J. (2023). Mathematical capabilities of ChatGPT. *arXiv* preprint, arXiv :2301 .13867.

Koehler, M. J., Shin, T. S., & Mishra, P. (2012). How do we measure TPACK? Let me count the ways. In R. N. Ronau, C. R. Rakes, & M. L. Niess (Eds.), Educational technology, teacher knowledge, and classroom impact: A research handbook on frameworks and approaches (pp. 16-31). IGI Global.

Mishra, P., & Koehler, M.J. (2006). Technological pedagogical content knowledge: A framework for integrating technology in teacher knowledge. Teachers College Record, 108(6), 1017-1054.

Patton, M.Q. (1990). *Qualitative Evaluation and Research Methods*, (2nd ed.), Newbury Park, CA: Sage

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15, 4–14.

Scherer, R., Siddiq, F., & Tondeur, J. (2019). The technology acceptance model (TAM): A meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. *Computers & Education*, 128, 13-35.

Assessment by skills of an interdisciplinary project in which mathematics is the common thread.

Maria Antonietta Lepellere[1] and Marzia Toso[2]

[1] University of Udine, Italy; maria.lepellere@uniud.it

[2]ISIS "Arturo Malignani", Udine; marzia.toso@malignani.ud.it

In this contribution we propose the design of an interdisciplinary STEAM project named "Dark ages?". It is based on Project Based Learning (PBL) and outdoor learning, in which digital and manipulative tools, virtual and real experiences, in the classroom and in the real world, are intertwined. A formative and summative evaluation proposal is presented, and the first results analysed. Some results of a survey on how students experienced the project are also presented.

Keywords: Project based learning, outdoor learning, STEAM, formative and summative assessment, survey, secondary education.

Introduction and Theoretical framework

In many contexts it is stated that one of the tasks of the school is to "educate complexity", train and consolidate the skills of reading reality as a complex system, where many variables operate. Interdisciplinarity can be a way to provide the student with overviews of complexity, in which points of observation, languages and interpretations intersect. From this perspective an interdisciplinary STEAM project "Dark ages?" was proposed. The title itself "Dark ages?" is an example of an interpretation called into question by the question mark. Who says that the Middle Ages are a dark, backward, and difficult period in European history? When? For what reason? Is that true? The question mark raises a doubt, asks questions, broadens perspectives, goes beyond the cliché. The aim of the interdisciplinary project is to bring students to the heart of an era, the medieval one, through STEAM and other subjects, as literature, history and English literature. Part of the project took place at school and another directly in the places studied: Siena Cathedral with the Museum and The Piccolomini bookshop and Pisa with the "Scuola Normale Superiore", the tower of the clock palace where Count Ugolino had been locked up and the cemetery where the statue of Fibonacci is. The following methodologies are used in the activities: Debate, Active Learning, Outdoor Education, Peer Tutoring, PBL. Furthermore, online research activities and evaluation of digital sources and resources, laboratory activities related to the creation/sharing of materials in the Moodle institutional e-learning area and in other shared digital areas as Genial.ly for the digital Escape Room and Padlet.com, were used in an active learning perspective. The technological tools used were: GeoGebra, Desmos, 3D print, Fusion 360° software, Teodolite for on-site measurements.

The term STEM refers to teaching and learning in the fields of science, technology, engineering, and mathematics. STEM education aims to help the next generation of students to solve real-world problems by utilizing knowledge of multiple disciplines and horizontal competences such as critical thinking, collaboration, and creativity. The addition of the artistic skills to the science and technology education gave birth to a new acronym: STEAM (notice the addition of A for arts). Zemelman, Daniels, and Hyde (2005) provide insight into the ten best STEM pedagogical practices for successful integration of STEM disciplines. There are as follows; using manipulatives and hands-on learning;

cooperative learning; discussion and inquiry; questioning and conjectures; (5) using the justification of thinking; writing for reflection and problem solving; using a problem-solving approach; integrating technology; teacher as a facilitator approach; using assessment as a part of instruction.

PBL has a lot of potential to enhance 21st century skills and engage students in real-world tasks (e.g., Kingston, 2018). It promotes interconnected worldview, links among disciplines, and presents an expanded view of subject matter (Blumenfeld et al., 1991; Kingston, 2018). Therefore, PBL is a promising teaching method for integrated science education that can be defined as an effort to organize or integrate science curriculum content into a meaningful whole by a constructive and context-based approach that crosses subject boundaries and links learning to real world (Czerniak & Johnson, 2014).

Learning outside the classroom essentially can be defined as use of resources out of the classroom to achieve the goals and objectives of learning (Knapp, 2010). The constant focus on textbooks and formal mathematical practice might invoke a view among students that mathematics is abstract, distanced and only useful in a in classroom context working only in the textbook. Existing research on outdoor learning in mathematics indicates positive affective outcomes and possible academic benefits from learning mathematics in an out-of-school context (Moffett, 2011). Moreover, outdoor environments, are real-life contexts enabling students to internalise, transfer and apply mathematical ideas and provides direct experience, and the students need to be active in the learning process. It lends itself to the Inquiry-based mathematics education a student-centred form of teaching whose guiding principle is that the students are supposed to work in ways like how professional mathematicians work (Artigue & Blomhøj, 2013). Lee et al. also suggested community partnerships, where students collaborate with professionals, as an important component of project based learning.

During the process of teaching and learning, teachers do assessment for learning (formative assessment) and assessment of learning (summative assessment). Black and Wiliam (2009) after considering the main features of teaching and learning defined formative assessment as: “Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited” (see also Cusi et al. 2017 and Aldon et al. 2017). Inquiry based learning needs formative assesment due to it process character.

In this contribution we propose the design an interdisciplinary STEAM project such that students are actors of their own knowledge through collaboration and discovery. The research questions are the following: How can this type of activity be assessed? How do students perceive this type of activity and the way to evaluate it?

Methodology and participants

The project was addressed to two 11th grade classes, at Applied Sciences High School of in Udine, Italy. In this article only one class of 19 pupils will be examined. The methodology used for the formative assessment was that of continuous monitoring, with related feedback, of the Forums, Padlets, Logbooks, delivery of the results of the proposed exercises and control via the Moodle eLearning platform and evaluation of the oral presentation of the works. For the final skills for summative assessment, we proposed a rubric containing six macro areas summarized by the following disjoint descriptors: Collaborate, Evaluate, Read and Understand, Communicate, Digital skills, Learn. A survey was also proposed at the end of the project to evaluate the appreciation and the effectiveness of the proposed project. In addition to a descriptive statistical analysis of the closed questions, a thematic analysis of the open questions was carried out. In the future, a comparative analysis of the results obtained will be carried out.

The design of the Project and the formative assessment

Interdisciplinary STEAM project requires a completely different way of grading. The normal teaching activity is also remodelled and reorganized, alternating laboratory/experiential moments with moments of study/empowerment. A working methodology for "open classes" is adopted: in some lesson hours, the classes are merged and reorganized into groups based on the proposed activity. The underlying theme of the project was to analyse a historical era, the Middle Ages, from various aspects to answer the question: was it really a "dark age?" Among the various parts of the STEAM project, we will describe those where mathematics plays a predominant role.

Activity 1. "Fibonacci: the man who gave us numbers". The students discovered unusual or hidden aspects of medieval (and non-medieval) mathematics through the resolution of a digital Escape Room and shared among peers the mathematical properties discovered. Before this activity the students only knew the definition of sequences as particular functions defined on the set of natural numbers. They had already studied the Fibonacci sequence in the previous school year and this year, with the IT teacher, they reviewed it together with the recurrence sequences. Among others, in the digital Escape Room there were also puzzles related to real life contest. As formative assessment the students, divided into groups, had to discover the solutions to the questions asked, using all the resources they deemed useful. They had to note down through logbooks the solutions of the questions, with the strategy used, and the difficulties encountered. The results were shared with the other groups in the classroom during the lesson. The module on "Fibonacci" concluded with the creation of a physical Escape Room, revisiting of Dante's Inferno, which allowed other classes of the Institute to "taste" the medieval world. It was entirely designed by the students including the divisions of the rooms and the interdisciplinary puzzles to move from one room to another, included the costumes and sets just like real directors and actors. They used the Moodle Forum for sharing ideas.

Figure 1: Digital and physical escape room. The Fibonacci statue.

Activity 2. "The white cathedral of Siena": a masterpiece of gothic architecture. Students, divided into groups, were asked to study the Cathedral of Siena from different point of view.

a) From a mathematical point of view: the students studied and (re)built some typical elements using the dynamic geometry software GeoGebra, such as friezes, rose windows, golden proportions, conics (see Figure 2 and 3). Individual student products were entered for evaluation in a Padlet.

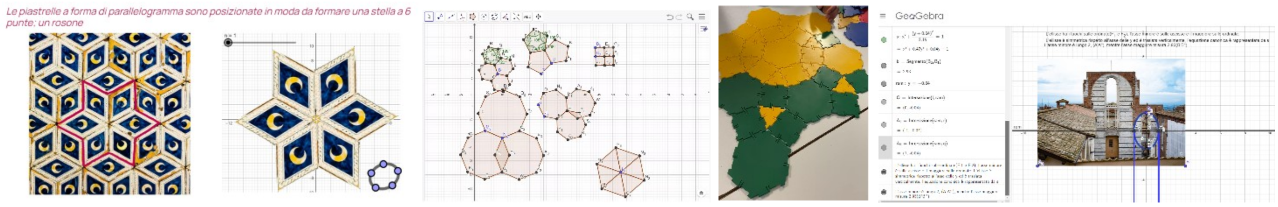
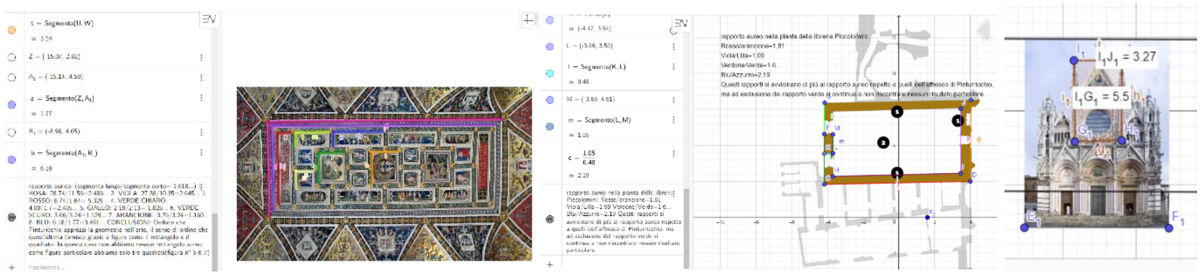

Le piastrelle a forma di parallelogramma sono posizionate in moda da formare una stella a 6 punte; un rosone
GeoGebra

Figure 2: Tessellation of a detail of cathedral with GeoGebra and manipulative objects. Find conics.


t = Segmento(U, W)
= 3.24
Z = (15.07, 2.82)
A_1 = (15.14, 4.59)
a = Segmento(Z, A_1)
= 1.77
B_1 = (-8.96, 4.65)
b = Segmento(A_1, B_1)
= 6.18

rapporto aurico: (segmento lungo/segmento corto= 1.618...) 1. ROSA: 28.74/11.56=2.486... 2. VIOLA: 27.38/10.35=2.645... 3. ROSSO: 8.74/1.64= 5.329... 4. VERDE CHIARO: 4.09/1.7=2.406... 5. GIALLO: 2.19/2.13= 1.028... 6. VERDE SCURO: 3.66/3.24=1.129... 7. ARANCIONE: 3.76/3.24=1.160... 8. BLU: 6.18/1.77=3.491... CONCLUSIONI: Deduco che Pinturicchio apprezzi la geometria nell'arte, il senso di ordine che quest'ultima fornisce grazie a figure come il rettangolo e il quadrato. In questo caso non abbiamo nessun rettangolo aureo, come figura particolare abbiamo solo tre quadrati(figura n° 5 6 7)

Inserimento...

= (-4.17, 3.54)
L = (-3.69, 3.56)
l = Segmento(K, L)
= 0.48
M = (-3.69, 4.61)
m = Segmento(L, M)
= 1.05
c = $\frac{1.05}{0.48}$
= 2.19

rapporto aureo nella pianta della libreria Piccolomini: Rosso/arancione=1.81 Viola/Lilla=1.99 Verdone/Verde=1.6... Blu/Azzurro=2.19 Questi rapporti si avvicinano di più al rapporto aureo rispetto a quelli dell'affresco di Pinturicchio, ma ad esclusione del rapporto verde si continua a non riscontrare nessun risultato particolare.

rapporto aureo nella pianta della libreria Piccolomini:
Rosso/arancione=1.81
Viola/Lilla=1.99
Verdone/Verde=1.6...
Blu/Azzurro=2.19
Questi rapporti si avvicinano di più al rapporto aureo rispetto a quelli dell'affresco di Pinturicchio, ma ad esclusione del rapporto verde si continua a non riscontrare nessun risultato particolare.

Figure 3: Let's search for the golden ratio!

b) From a historical-philosophical-architectural point of view, they studied the floor "Come stelle in terra" analysing the proposed allegorical meanings. In Figure 4 there are some examples.

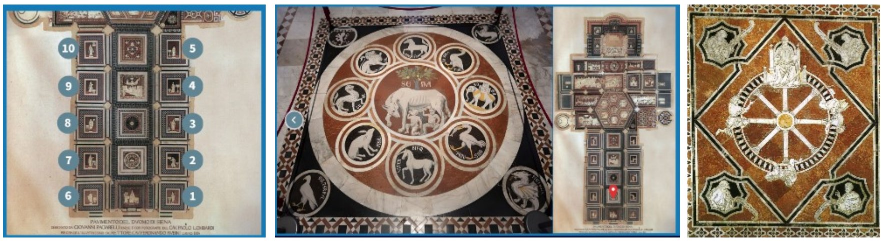


Figure 4: Allegorical part

c) From Technology and Engineering point of view, with Peer tutoring, they tried to 3D print some elements by developing the Fusion 360° software and learned to use the Teodolite.

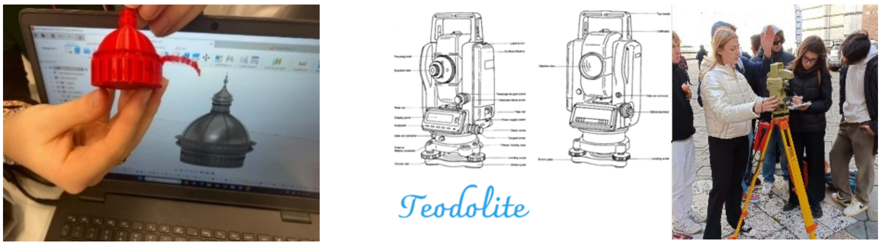


Figure 5: 3D print and Teodolite

As final activities of the project they also studied the Unity language to create a virtual museum, using 360-images. In figure 6 there are some examples.

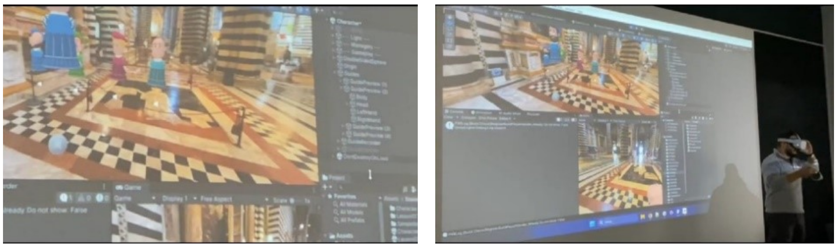


Figure 6: Multimedia production

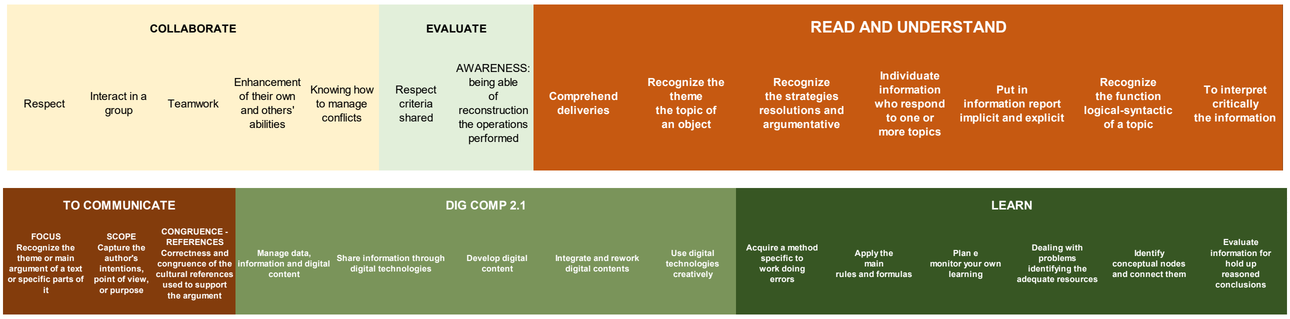
Activity 3. The outdoor activity: The visit to Pisa and Siena. The visit included laboratories as: Physical measurements with the laser and with Teodolite, as an application of the trigonometry, using the plans provided to us directly by the Opera di Siena; mathematical and architectural studies with related data analysis; take photos, videos with the aim of preparing a virtual museum.

The students had the opportunity to delve deeper into the topics covered thanks to the interventions of experts: The architect De Benedetti of the of Opera for an architectural, geophysical, and historical-cultural point of view; Professor Bellissima, with a contribution of music to mathematics, who discussed Guido d'Arezzo and, starting from the miniatures contained in the Piccolomini Chapel which show the first forms of musical writing, he led the students to play with functions on the Cartesian plane which represent the various forms of musical writing from the beginning up to digital music; Professor Chiantini, who discussed "The fields of fortune". He starts from a marble mosaic in the Cathedral of Siena, called "The wheel of fortune", and proposed a reflection on medieval cosmology in the light of non-Euclidean geometries.

From formative to summative assessment

In relation to the final skills achieved by the students, we propose a rubric, inspired to The Periodic Table of Skill, divided into 6 macro areas: Collaborate, Evaluate, Read and understand, Communicate, Digital skills (Dig Comp 2.1), Learn. The group of teachers has chosen, under the guidance of the researcher, inspired to the periodic table (Table 1), to use for each macro competence, some items which are best suited to representing all the disciplines involved. Collaborate includes respect, group interaction, teamwork, valorisation of one's own and others' abilities, knowing how to manage conflicts. Evaluate includes respect shared criteria, to be able to reconstruct the operations carried out. Read and Understand include understanding the deliveries, recognizing the topic, recognize resolution and argumentative strategies, identify information that responds to one or more topics, relating implicit and explicit information, recognize the logical-syntactic function of an argument, critically interpret information. Communicate include, recognize the main theme or argument of a text or specific parts of it, capture the intentions and the author's point of view, correct the cultural references used to support the argument. Digital skills include manage data, information and digital content, share information through digital technologies, develop digital content, integrate and rework digital content, use digital technologies creatively. Learn include acquiring a specific working method by making mistakes, apply the main rules and formulas, plan and monitor learning, deal with problems by identifying the appropriate resources, identifying conceptual nodes and connecting them, evaluating information to support reasoned conclusions.

Table 1: The Periodic Table of Skills



COLLABORATE					EVALUATE		READ AND UNDERSTAND						
Respect	Interact in a group	Teamwork	Enhancement of their own and others' abilities	Knowing how to manage conflicts	Respect criteria shared	AWARENESS: being able of reconstruction the operations performed	Comprehend deliveries	Recognize the theme the topic of an object	Recognize the strategies resolutions and argumentative	Individuate information who respond to one or more topics	Put in information report implicit and explicit	Recognize the function logical-syntactic of a topic	To interpret critically the information

TO COMMUNICATE			DIG COMP 2.1					LEARN					
FOCUS Recognize the theme or main argument of a text or specific parts of it	SCOPE Capture the author's intentions, point of view, or purpose	CONGRUENCE - REFERENCES Correctness and congruence of the cultural references used to support the argument	Manage data, information and digital content	Share information through digital technologies	Develop digital content	Integrate and rework digital contents	Use digital technologies creatively	Acquire a method specific to work doing errors	Apply the main rules and formulas	Plan e monitor your own learning	Dealing with problems identifying the adequate resources	Identify conceptual nodes and connect them	Evaluate information for hold up reasoned conclusions

The need to have 2 votes on the register led us to unify some macro-sectors. In the first we merged Collaboration, profitable relationships, respect, civil discussion of ideas, management of other people's places and objects, theatrical interpretation and problem solving. The second included: interdisciplinary path, mathematical study, argumentative and expository ability (Table 2.)

Table 2: Student results

<table>
<tr><th colspan="19">Collaboration, profitable relationships, respect, civil discussion of ideas, management of other people's places and objects, theatrical interpretation and problem solving.</th></tr>
<tr><td>10</td><td>8,5</td><td>9,5</td><td>9,5</td><td>7,5</td><td>9</td><td>8,5</td><td>9</td><td>10</td><td>10</td><td>9,5</td><td>9</td><td>10</td><td>9</td><td>9</td><td>9</td><td>10</td><td>8,5</td><td>8,5</td></tr>
</table>

<table>
<tr><th colspan="19">Interdisciplinary path, mathematical study, argumentative and expository ability</th></tr>
<tr><td>10 -</td><td>7</td><td>10</td><td>8</td><td>7</td><td>9</td><td>8</td><td>9-</td><td>8</td><td>8</td><td>10-</td><td>7</td><td>9</td><td>7</td><td>10</td><td>8</td><td>8</td><td>8</td><td>7</td></tr>
</table>

As can be seen, the students' grades, from 0 to 10, were very high thanks to the interdisciplinarity of the project with a mean of 8 in the first case and 9 in the second. In 8 cases out of 19 the scores for the two types of judgment were close to each other (with a difference of 1 or 2 point). The scores of the first evaluation framework were far higher than the second in 10 cases over 19 while just in 1 case the second type of evaluation gave far better result. There is a positive correlation 0.41 between the two evaluations, meaning that the two ratings are not independent in the sense that the second is positively influenced by the first. But this correlation is not very high, which means that the two assessments managed to capture specific skills that would not have emerged with a single assessment.

The survey

A project evaluation survey was also proposed to the students. We report the results only regarding 4 over 36 questions: 1. Which aspect did you appreciate the most? 2. Which aspect bothered you the most? 3. I think that evaluating an interdisciplinary activity is adequate and correct. 4. How likely are you to recommend this activity to another class council at your school?

1. Which aspect did you appreciate the most? The appreciation of the group work and with the other classes and the visit to Siena were the most used. We report just some sentences in this direction: I appreciate the concrete observation of what was studied during the visit to Siena; Moments of discussion, brainstorming, research, and related sharing. The division of the material to be studied between the classes and between the groups created in the classes. This allows us to delve into a specific topic in an exhaustive way without being totally unaware of the other aspects related to the macro topic; I appreciate the exchange of information between classes, especially when one class exposed something to the other class, the class listening could intervene by adding content or asking questions.

2. Which aspect bothered you the most? Most students did not find any negative aspects, Just for few the main problem is the time: required for the project seemed excessive for the fear of having to tackle the program too quickly, but not enough for visiting the cities.

3. I think that evaluating an interdisciplinary activity is adequate and correct. The possibility for the answers were: strongly disagree, disagree, neutral, agree and strongly agree. The 42% were strongly agree, the 26% were agree, the 32% were neutral and nobody disagree or strongly disagree.

4. How likely are you to recommend this activity to another class council at your school? On a scale from 0 to 10, only two gave a score of 4, the other was greater than 7 with a total mean of 8.

The project was appreciated, and this serves as an encouragement to design new ones.

Conclusions

The experience of an interdisciplinary activity where different student skills is brought into play is a challenge for both teachers and stimulating students. PBL involves a dynamic classroom approach, which emphasizes on long-term learning, interdisciplinary and student-centered activities. Students need both manipulative and technological objects and need to experience on-site activities. The possibility of interacting with teachers other than one's own and with experts is an opportunity to compare various communication and operational methods and offers opportunities for discussion, socialization, and integration between students. The activity also facilitates dialogue and collaboration between teachers, encouraging the sharing of effective strategies.

Students' assessment should be considered an integral part of instruction. Each instructional activity could be seen as an opportunity for the teacher to assess as well as for students to learn. Emphasis should be on formative assessment that aims at supporting students learning. This includes reflection, self and peer evaluation, and teachers' feedback throughout the project process. Assessment should include a specific end-of-project phase that ensures reflection on what was learned as well as the creation of a project artefact. A public presentation of the project supports students' communication skills, can motivate students, and presents an opportunity for feedback. Instead of a presentation, the product itself can be public.

We believe that A (for arts and design) in STEAM is an important addition from the original STEM approach. Arts and design permit a more divergent thinking in students, giving space to more creative solutions to problems. For a real implementation of STEAM, it is needed a radical change in the educational culture and administration. Curriculum and teacher assessment methods should change giving importance not only to the concepts but also the acquisition of transversal competences. Our findings highlight the importance of having an assessment rubric with criteria that addressed both disciplinary specific skills as well as generic STEM skills.

The excellent results obtained encourage us to promote other projects like this. But evaluating this type of project is not easy and it is a challenge that teachers are called upon to accept to have students who are increasingly ready to face the complexity of our times.

Acknowledgment

This project was possible thanks to the precious collaboration of the teachers of the "Malignani" school in Udine Brocato Maria Concetta, Donnini Letizia and Fabris Letizia. We also thank the professors Zangrando Nicola of Udine University and Bellissima Fabio, Chiantini Luca, Scala Riccardo and Scianna Giuseppe of Pisa University.

References

Aldon, G., Cusi, A., Morselli, F., Panero, M., & Sabena, C. (2017). Formative assessment and technology: reflections developed through the collaboration between teachers and researchers. In

G. Aldon, F. Hitt, L. Bazzini & U. Gellert, *Mathematics and technology: a CIEAEM source book. Series 'Advances in Mathematics Education'*. Springer International Publishing.

Artigue, M., & Blomhøj, M. (2013). Conceptualizing inquiry-based education in mathematics. *ZDM*, *45*(6), 797-810.

Black, P., & Wiliams, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21, 5–31.

Blumenfeld, P. C., Soloway, E., Marx, R. W., Krajcik, J. S., Guzdial, M., & Palincsar, A. (1991). Motivating project-based learning: Sustaining the doing, supporting the learning. *Educational Psychologist, 26*(3-4), 369–398.

Cusi, A., Morselli, F., & Sabena, C. (2017). Promoting formative assessment in a connected classroom environment: design and implementation of digital resources. *ZDM, 49*, 755-767.

Czerniak, C. M., & Johnson, C. C. (2014). Interdisciplinary science teaching. In S. K. Abell, & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 395-411)

Kingston, S. (2018). Project based learning & student achievement: What does the research tell us? *PBL Evidence Matters*, 1(1), 1-11.

Knapp, C. E. (2010). Teaching for Experiential Learning: Five Approaches That Work. *Journal of Experiential Education*, 33, 288-290.

Lee, J. S., Blackwell, S., Drake, J., & Moran, K. A. (2014). Taking a leap of faith: Redefining teaching and learning in higher education through project-based learning. *Interdisciplinary Journal of Problem-Based Learning, 8*(2), 2. https://doi.org/10.7771/1541-5015.1426

Moffett, P. V. (2011). Outdoor mathematics trails: an evaluation of one training partnership. *Education* 3-13, 39(3), 277-287.

Zemelman, S., Daniels, H., & Hyde, A. (2005). Best practice: New standards for teaching and learning in America's school (3rd Edition). Portsmouth, NH: Heinemann

Advancing assessment in fractions: designing and implementing formative proficiency tasks

Hui-Chuan Li

University of Edinburgh, Moray House School of Education and Sport, Edinburgh, United Kingdom; huichuan.li@ed.ac.uk

While many studies have focused on the challenges students face in the domain of fractions and the corresponding pedagogies of teaching and learning, there has been inadequate and disproportionate attention dedicated to assessment resources, particularly those tailored for formative assessment in the context of fractions. This study takes a step forward in contributing new insights to this field by designing fraction proficiency tasks explicitly intended for formative assessment of students' comprehension of fractions. These fraction proficiency tasks were administered to a class of 35 fifth-grade students (ages 10–11) with mixed abilities in a primary school in Taiwan to evaluate their understanding of fractions. Findings of the study offer valuable insights into assessing students' understanding of fractions and provide a comprehensive view of the diversity in students' understanding and the extent of these differences. Implications for future studies are also presented.

Keywords: Fractions, fraction proficiency, assessment resources, assessment for learning, formative assessment.

Introduction

The teaching and learning of fractions persistently present challenges for both teachers and students. It has been argued that, for many students, learning fractions often involves merely manipulating symbols to arrive at the correct answer. While they might employ the appropriate fractional terms and solve some fraction-related problems, several critical aspects of fractions still escape them (Soni & Okamoto, 2020). Students' struggles with fractions often stem from the intricate relationships between various representations and fundamental arithmetic operations (Cramer et al., 2002), wherein the simultaneous symbolic nature of fractions contributes to these challenges. For mathematicians, fractions are rational numbers expressible in the form "a / b" where b ≠ 0, rather than simply representing parts of wholes. They are not just ratios of two natural numbers but also constitute numbers in themselves.

In school, many children often receive only brief exposure to the concepts and procedures of fractions and are taught fraction algorithms with minimal emphasis on their conceptual underpinnings (Lenz et al., 2022). One of the conventional concrete approaches to learning about fractions often involves thinking in terms of partitioning or equal sharing. It is wise to base this idea on discrete countable objects, as well as on objects that may require the dissection of a continuous whole. However, this concrete approach does not cover the entirety of understanding fractions. For instance, the concept of "equal sharing" is just one among many properties of fractions and, on its own, is not adequate to convey a meaningful understanding of fractions to children.

Much research has concentrated on the challenges students encounter specifically within the field of fractions and their associated teaching and learning pedagogies, However, there has been insufficient and disproportionate attention given to assessment resources, especially for formative assessment, specifically designed for fractions. The significance of formative assessment, also referred to as assessment for learning, lies in its capacity to offer continuous feedback and insights into students' understanding and progress. This study has taken a step forward in contributing new insights to this field by designing fraction proficiency tasks explicitly intended for formative assessment of students' comprehension of fractions. In doing so, it seeks to provide valuable insights into the assessment of students' understanding of fractions, offering a comprehensive perspective on the diversity and extent of challenges encountered by students.

Fraction proficiency tasks

Drawing from Tsai and Li's (2017) fraction proficiency framework and an extensive review of fraction-related studies, 15 tasks were designed to assess students' comprehension levels and identify areas where they might encounter difficulties in the field of fractions. The content of these 15 tasks was specifically organized in four major topics to encompass the five dimensions of fraction proficiency identified by Tsai and Li (2017), which include: (1) the part-whole, measure, quotient, operator and ratio constructs of fractions, (2) the concept of equivalent fractions, (3) the procedural fluency for and conceptual understanding of fraction operations, (4) the relationship between fractions, decimals and percentages, and (5) the transition between different forms of representations involving fractions. Lesh's (1981) representation model (Dimension 5) was integral to all tasks that required students to solve problems by transitioning between representations. In the following sections, examples of tasks for each topic will be provided.

Topic 1: Five constructs of fractions

Topic 1, centered on five constructs of fractions (Dimension 1), involved designing four tasks aimed at assessing students' understanding of part-whole, measure, quotient, operator, and ratio constructs related to fractions. The statements for these four tasks are provided in Table 1.

Topic 2: Equivalent fractions

In Topic 2, focusing on equivalent fractions (Dimension 2), three tasks were formulated to explore students' comprehension of equivalent fractions and their conceptualization of expanding and reducing fractions to determine an equivalent form. Table 2 outlines these three tasks.

Topic 3: Multiplication of fractions

Topic 3, emphasizing the multiplication of fractions (Dimension 3), comprised five tasks designed to assess students' understanding of the reasoning behind their procedural skills in performing fraction multiplication. The statements for these four tasks are provided in Table 3.

Topic 4: Fractions, decimals and percentages

In Topic 4, three tasks were devised to examine the extent to which students recognize the relationship between fractions, decimals, and percentages (Dimension 4).

Table 1: Descriptions for tasks related to Topic 1

	Statement of Task	Transitions between representations (Dimension 5)
Topic 1 Five constructs of fractions (Dimension 1)	What is a fraction? How would you would explain to someone what a fraction is? Please offer three different explanations, and one or more of your explanations needs to relate to a real-life situation.	From the fractional symbolic representation to the real-life situation representation.
	Are they reasonable to you? (Students are given a set of cards that visually represent a fraction) Please select the cards you consider reasonable and explain your reasoning behind your choices.	From the pictorial representation to the fractional symbolic representation.
	Who spends more? Mary and John went to McDonalds. Mary spends 1/4 of her pocket money and John spends 1/2 of his. Do you agree it is possible that Mary spent more than John? Why do you think this?	From the real-life situation representation to the fractional symbolic representation.
	Which ones are reasonable? (Students are given a set of cards that visually represent a fraction) Please look at the cards provided; which of these cards are reasonable and which are not?	From the pictorial representation to the fractional symbolic representation

Table 2: Descriptions for tasks related to Topic 2

	Statement of Task	Transitions between representations (Dimension 5)
Topic 2 Equivalent fractions (Dimension 2)	What are equivalent fractions? (students are given a set of cards that visually represent an equivalent fraction). Please write down your observations from these cards, and then elaborate on how your findings are connected to equivalent fractions.	From the pictorial representation to the fractional symbolic representation
	Match the pairs (students are given a set of cards that visually represent an equivalent fraction) Please match equivalent fractions from these cards and then explain how you paired them.	From the pictorial representation to the fractional symbolic representation
	Who gets more? At two different tables where 2 children were sharing 3 chocolate bars and 6 children were sharing 9 chocolate bars. These chocolate bars are all the same size. Please indicate who will receive more and elaborate on your thought process behind your choice.	From the real-life situation to the spoken representation

Table 3: Descriptions for tasks related to Topic 3

	Statement of Task	Transitions between representations (Dimension 5)
Topic 3 Multiplication of fractions (Dimension 3)	Let's fold a paper fraction (Students are given pieces of colour paper) Please fold the fractions: 1/8, 1/6 and 1/12 using the paper provided.	From the fraction symbolic representation to the manipulative representation
	Jenny's birthday party Jenny wants to invite her three best friends to come to her party. Each of her three friends can consume 3/4 of a pizza Please illustrate with a diagram to show the quantity of pizza Jenny will require, and then provide a mathematical written representation to represent your drawing.	From the real-world situation to the pictorial representation
	What do you think 2/3 × 5? Please provide a real-life scenario that represents the mathematical operation 2/3 × 5 and then use a drawing to represent 2/3 × 5.	From the fractional symbolic representation to the real-life situation and to the pictorial representation
	How much cake had Jenny's brother eaten? Jenny's mum made a square-shaped cake for her birthday. At the party, half of the cake was eaten and then the rest was put in fridge. The next day, Jenny's brother ate 2/3 of the remaining part of the cake. (Students are given pieces of color paper) How would you fold the paper to illustrate the portion of cake Jenny's brother had consumed? Afterwards, provide a written mathematical representation to explain the folding method.	From the real-life representation to the manipulative representation
	What do you think 1/4× 3/4? Please provide a real-life scenario that represents the mathematical operation 1/4× 3/4 and then use a drawing to represent 1/4× 3/4	From the fractional symbolic representation to the real-world representation and to the pictorial presentation

Data collection and analysis

This study recruited a class of 35 fifth-grade students (ages 10–11) with mixed abilities from a primary school in Taiwan to evaluate their understanding of fractions. The four topics covered in the fraction proficiency tasks had been introduced to the participants in their previous school years as part of the current mathematics curriculum in Taiwan. The fraction proficiency tasks were given to all students without time limits for completion. Most students finished within 60 to 70 minutes and submitted their answer sheets and materials (such as cards and colored folding paper) to their teacher upon task completion.

Table 4: Descriptions for tasks related to Topic 4

	Statement of Task	Transitions between representations (Dimension 5)
Topic 4 Fractions, decimals and percentages (Dimension 4)	What is a percentage? Please explain what a percentage is and provide some examples from your life where percentages are commonly observed.	From the fractional symbolic representation to the real-life situation representation
	To what extent is Tom sure? When Tom is going to school, his mum asks him if he is prepared well for his school test today. Tom replies: “Yes”. Mum asks: “Are you sure?”, Tom says: “One hundred per cent sure”. Mum asks: “So you will get a full mark home, will you?” Tom makes a funny face and says: “Um, um, fifty per cent sure”. Please explain the meanings of “one hundred per cent sure” and “fifty per cent sure,” and then demonstrate how fractions can represent these expressions.	From the real-life situation representation to the fractional symbolic representation
	How are they related? Here are three numbers: 0.4, 2/5 and 40%. Please explain the relationship between these three numbers and how they can be converted from one form to another.	From the symbolic representation to the spoken representation

Apart from the data collected from the students’ drawings and paper folding exercises, which were neither numerical nor narrative, much of the data collected was in the form of words. An inductive coding approach was employed to identify both general and distinctive features from the texts, following these three steps: identifying and labeling, reducing, and summarizing. This aligns with Thomas’s (2006) assertion that inductive approaches are designed to facilitate an understanding of meaning in complex data by developing summary themes or categories derived from the raw data.

The nature of students’ fraction understanding

What can be learned from Topic 1?

This topic, based on Kieren’s (1988) theory, assessed students’ comprehension of the five constructs of fractions, revealing that their understanding of fractions was either confused or incomplete. Their grasp of fractions predominantly revolved around the part-whole construct, with minimal consideration for the equality of each part of the whole. This aligns with existing literature suggesting an excessive focus on the part-whole construct, hindering students' ability to position fractions on a number line (Saxe et al., 2013). Moreover, the challenge students faced in positioning 3/5 on a number line in this study further confirms the findings of Soni and Okamoto (2020), emphasizing students’ struggles in locating fractions accurately on a number line. Another challenge observed was the students’ inflexible recognition of a fraction’s unit. Chan et al. (2007, p. 26) also argued that “the units concept is a common conceptual deficiency among students, indicating a significant flaw in current fraction teaching practices in Taiwan”.

What can be learned from Topic 2?

An analysis of students' responses to the tasks in Topic 2 supported earlier research findings (Lamon, 2007) that students' reasoning of equivalent fractions was rather rule-based. For example, the "Who gets more" task in Table 2 showed that 20 out of 35 students answered it correctly. However, their explanations generally referred to "Because 3/2=9/6"; "Because they are equivalent fractions" or "By using the rule of expansion or reduction, you then know they are the same" to explain how they solved the task. It is not wrong to describe the equivalence of fractions based on the rules of expansion or reduction, but there is a danger that students apply "rule-based" explanations without understanding them (Levenson et al., 2004). This rule-based emphasis is also echoed in Yang's (2005) finding that both teachers and students tended to "rely on rule-based methods to explain their reasoning" in the field of fractions. This suggests the importance of allowing students the opportunity to articulate their thoughts and construct explanations that are not solely rule-based.

What can be learned from Topic 3?

An understanding of fraction multiplication often challenges students because they have to distinguish it from multiplication of whole numbers, that is, from repeated addition to multiplicative reasoning. An operator construct of fractions is fundamental for interpreting the meaning behind the multiplication of fractions (Thompson & Saldanha, 2003). In this topic, the "What do you think 2/3 × 5?" task (see Table 3) shows that multiplication of a whole number and a proper fraction was presented by more than half of the students as repeated addition (e.g., 2/3 × 5= 2/3 + 2/3 + 2/3 + 2/3 + 2/3). Such additive reasoning, although it provides a useful connection between multiplication and addition, may not be meaningfully interpreted for multiplying two proper fractions, which would produce a smaller fraction. This may explain why over half of students encountered challenges in providing a real-life scenario to represent the mathematical operation 1/4× 3/4 and to depict 1/4 × 3/4 through a drawing when responding to the "What do you think 1/4× 3/4?" task in this topic.

What can be learned from Topic 4?

In this topic, students' responses to the "What is a percentage?" task (see Table 4) highlighted their struggles in articulating their reasoning behind percentages. However, their responses to the "How are they related?" task revealed that 28 out of 35 students were capable of converting procedurally between these three different forms. This suggests that the students in this study recognised the quotient construct of a fraction – i.e. 2/5 means 2÷ 5 and 0.4 means 4 ÷10 – and, when dividing the numerator by the denominator, they had no problem converting from a fraction to a decimal. This proficiency contrasts with Moss's (2005) findings, where over half of the students (sixth and eighth graders in Canada) claimed that "1/8 would be 0.8" when expressed as a decimal. This emphasizes the critical role of the quotient construct in comprehending the connection between fractions and decimals. Moreover, the outcomes of this topic demonstrated students' proficiency in employing various strategies to convert between fractions, decimals, and percentages, suggesting they understood the relationships between three different forms that have identical values.

Fraction proficiency tasks for formative assessment of students' comprehension of fractions

Formative assessment takes various forms, aiding both students and teachers in evaluating learning objectives and adjusting instruction. Fraction proficiency tasks, as demonstrated earlier, can serve as formative assessments, enabling students to demonstrate their skills and identify errors and misconceptions in fractions. Within the classroom, these tasks can seamlessly integrate into ongoing formative assessment practices, allowing teachers to gain insights into student progress, deepen understanding of fractions, and address individual learning needs efficiently. Teachers observe student engagement, provide immediate feedback, and encourage self-assessment, fostering metacognitive skills and ownership of learning. Peer assessment can further enhance learning by providing diverse perspectives and collaborative feedback (Black et al., 2003).

Limitations, implications and directions for future research

This study focused on a specific mathematical area – fractions – and it only examined students' understanding of fractions based on the Tsai and Li's (2017) framework. It is recognized that various other aspects pertaining to fractions might not have been incorporated in these tasks; also, other related factors might not have been taken into account. The findings are confined by the constraints of the employed methodology as well as the limitations inherent in the sample. However, the fraction proficiency tasks presented in this study offer valuable insights into assessing primary students' understanding of fractions. They provide a comprehensive view of the diversity in students' understanding and the extent of these differences.

This study shows that the difficulties encountered by students in the present study resonate with those identified in previous studies. This suggests that fractions continue to pose challenges for students, even among Taiwanese students who are consistently recognized as high-performing in large-scale mathematics comparative assessments such as TIMSS and PISA. The findings of this study also offer assessment resources for teachers to gain a clearer understanding of what students should attain and what areas they need to develop. This assists in integrating diverse aspects of fraction knowledge, aiding both students and teachers in comprehending fractions more effectively.

Another implication of this study for fraction assessment involves reconsidering the role of assessment in contributing to a broader understanding of students' grasp of fractions and their mathematical knowledge overall. As argued by Saxe et al. (2013), fractions-related topics are often seen as disconnected. Therefore, further research is needed, particularly in formative assessment, where evaluating a comprehensive understanding of fractions across multiple facets should take precedence over isolating one facet from the others.

References

Black, P., Harrison, C., Lee, C. Marshall, B., & Wiliam D. (2003) *Assessment for learning*, Maidenhead: Open University Press.

Chan, W.-H., Leu, Y.-C., & Chen, C.-M. (2007). Exploring group-wise conceptual deficiencies of fractions for fifth and sixth graders in Taiwan. *The Journal of Experimental Education*, *76*(1), 26–57. https://doi.org/10.3200/JEXE.76.1.26-58

Cramer, K. A., Post, T. R., & delMas, R. C. (2002). Initial fraction learning by fourth- and fifth-grade students: a comparison of the effects of using commercial curricula with the effects of using the rational number project curriculum. *Journal for Research in Mathematics Education*, *33*(2), 111–144.

Kieren, T. E. (1988). Personal knowledge of rational numbers: its intuitive and formal development. In J. Hiebert & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 162–181). National Council of Teachers of Mathematics.

Lenz, K., Reinhold, F., & Wittmann, G. (2022). Topic specificity of students' conceptual and procedural fraction knowledge and its impact on errors. *Research in Mathematics Education*, https://doi.org/10.1080/14794802.2022.2135132

Lesh, R. (1981). Applied mathematical problem solving. *Education Studies in Mathematics*, *12*, 235–264. https://doi.org/10.1007/BF00305624

Levenson, E., Tirosh, D., & Tsamir, P. (2004). Elementary school students' use of mathematically-based and practically-based explanations: the case of multiplication. *Proceedings of the 28th conference of the international group for the Psychology of Mathematics Education (PME)*, *3*, 241–248, https://eric.ed.gov/?id=ED489578

Moss, J. (2005). Pipes, tubes, and beakers: new approaches to teaching the rational-number system. In M. S. Donovan & J. D. Bransford (Eds.), *How students learn: history, mathematics, and science in the classroom* (pp. 309– 350). National Academies Press.

Saxe, G. B., Diakow, R., & Gearhart, M. (2013). Towards curricular coherence in integers and fractions: a study of the efficacy of a lesson sequence that uses the number line as the principal representational context. *ZDM-International Journal on Mathematics Education, 45*, 343–364. https://doi.org/10.1007/s11858-012-0466-2

Soni, M. & Okamoto, Y. (2020). Improving children's fraction understanding through the use of number lines. *Mathematical Thinking and Learning*, *22*(3), 233–43. https://doi.org/10.1080/10986065.2020.1709254

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2), 237– 46. https://doi.org/10.1177/1098214005283748

Thompson, P. W., & Saldanha, L. A. (2003). Fractions and multiplicative reasoning. In J. Kilpatrick, W. G. Martin & D. Schifter (Eds.), *A research companion to principles and standards in school mathematics* (pp. 95–113). National Council of Teachers of Mathematics.

Tsai, T.-L. & Li, H.-L, (2017). Towards a framework for developing students' fraction proficiency. *International Journal of Mathematical Education in Science and Technology*, *48*(2), 244– 255, https://doi.org/10.1080/0020739X.2016.1238520

Yang, D. C. (2005). Number sense strategies used by 6th-grade students in Taiwan. *Educational Studies, 31*(3), 317–333. https://doi.org/10.1080/03055690500236845

Elaborated feedback for interpreting graphs in Secondary Education

María Sanz-Ruiz[1], Luis Miguel Soto-Sánchez[1], Zaira Ortiz-Laso[2], José Manuel Diego-Mantecón[1]

[1]Universidad de Cantabria, Spain; sanzm@unican.es, luis-miguel.soto@alumnos.unican.es, diegojm@unican.es

[2]Universidad de Alicante, Spain; zaira.ortiz@ua.es

Interpreting and extracting information from graphs can be challenging for secondary education students. While feedback often yields positive effects in correcting student errors, there remains a gap in understanding how prior knowledge influences performance in the context of functions and their reactions to feedback. In this study, 68 students solved a task with a graph through an electronic assessment tool (STACK) that allows multiple opportunities to solve it and progressively provides feedback for arriving at the solution. Our results reveal that most medium achievers made a standard error and overcame it after receiving one or two hints, whereas low achievers committed uncommon mistakes and required at least three hints. Our study also shows different reactions to feedback; low achievers were likely to feel overwhelmed when connecting the feedback provided across multiple attempts as the information was presented individually.

Keywords: Electronic assessment, feedback, functions, graphs, STACK.

Graphs and feedback through e-assessment

Interpreting graphs is essential for secondary education students to tackle real-life situations and advance academically (Planinic et al., 2012). Research has revealed difficulties around this topic (Graham & Sharp, 1999; Ortiz-Laso, 2017), often concerning the identification of which graph features should be used to extract information (Graham & Sharp, 1999). In this line, Ruchniewicz and Barzel (2019) outlined that electronic assessment (e-assessment) tools help students reflect on graph interpretation and attain the knowledge needed to solve a mathematical task.

Task-related knowledge can be delivered through diverse feedback varying from simple to elaborated. Simple feedback relates to how well a task has been performed (Narciss et al., 2022) and comprises three types: knowledge of result (KR), knowledge of performance (KP), and knowledge of correct result (KCR; Narciss, 2008). KR provides information about response correctness (e.g., correct or incorrect), KP gives the number of correct responses, and KCR delivers the correct task solution. Elaborated feedback delivers concise information and can be divided into five types: knowledge on task constraints (KTC), knowledge about concepts (KC), knowledge about mistakes (KM), knowledge on how to proceed (KH), and knowledge on metacognition (KMC) (Narciss, 2008). KTC clarifies task nature, subtasks, processing rules, and requirements, for example "The first step of the correct solution would be…" (Pinkernell et al., 2020, p. 223). KC delivers conceptual information to reach the solution, such as offering a mathematical definition, whereas KM identifies mistakes' location, type and origins, for example "You probably made this error..." (Pinkernell et al., 2020, p. 223). KH guides the responder to the right solution, correcting specific mistakes and presenting hints and examples, like "Do not ignore the cards that are negative instances of the given concept, as they provide useful information" (Narciss, 2013, p. 19). Finally, KMC offers guiding questions attracting attention to metacognitive strategies (Narciss et al., 2022).

Previous studies on elaborated feedback have investigated its usefulness according to students' previous knowledge. Fyfe and Rittle-Johnson (2016) state that both high- and low-achievers benefit from elaborated feedback, whereas Pinkernell et al. (2020) claim that only low achievers improve. The type of elaborated feedback also influences its effectiveness; Pinkernell et al. (2020) show that among German low-achieving secondary education learners, the benefits derived from KM or KTC surpassed those obtained from KH.

Research focus and methods

To shed light on the effectiveness of elaborated feedback, this study investigated how secondary education students request and react to feedback provided through an e-assessment tool and how it influences attaining the correct solution. A sample of 68 students from two high schools in Cantabria, Spain, was selected. Students were in the third and fourth grades of compulsory secondary education, aged 14 to 16. They received instruction through textbooks on functions and graphs and represented three achieving groups: low, medium, and high.

Proposed task and designed feedback

To assess our objectives, we adopted one of the tasks of Ortiz-Laso's (2017) graph, illustrated in Figure 1. In her research, student responses were classified into three sets: correct answers (100 min), expected error (120 min; students did not realize that Juan was not moving between minutes 50-70), and unexpected errors (non-typical errors related to students' lack of knowledge on graphs).

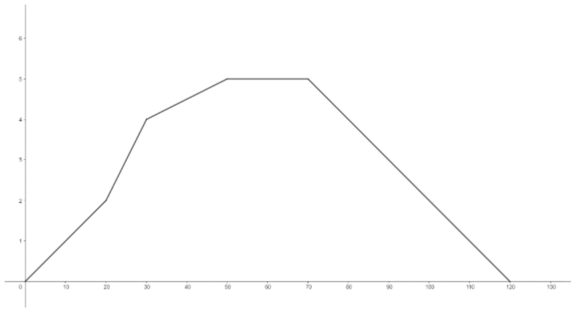


Figure 1. Juan's distance from home

Juan leaves home to exercise in a mountain zone. He starts walking at his usual pace and then alternates between running and walking at different paces. The graph in Figure represents his activity, where the x-axis is time (in minutes) and the y-axis is the distance from home (in kilometers). How much time does Juan spend in motion?

To deliver elaborated feedback on the errors classified by Ortiz-Laso, we considered three Narciss' (2008) categories (KR, KH, and KC) to be displayed through STACK (System for Teaching and Assessment using a Computer Algebra Kernel). This tutoring system was chosen because it dispenses specific feedback for each error and tracks every response (Sangwin, 2013). KR feedback was designed for correct answers, whereas KH and KC were devised for the wrong ones (Figure 2). For the expected error (120 min), the solvers received automatic KH to reflect on each graph part (A: *Is Juan not moving at any moment?*). Then, they also could request progressive hints as follows: extra KH (A1: *What does it mean for the graph to have a horizontal part?*), KC related to the constant part of the graph (A2: *The graph indicates the distance between Juan and his home. If the graph is constant during a period, it means that Juan is not moving during that time.*), and further KC incorporating an explanation about the time variable (A3: *Remember that the walk lasted for 120 minutes, and Juan was not moving between minutes 50-70*). For unexpected errors, the solvers got automatic KC that included a graph description (B: *The graph shows the distance between Juan and* *his home at each moment. His route finishes when the distance is 0 again*). Finally, they also had the opportunity to request KC and KH to get a task reformulation (B1: *The graph shows Juan was outside for 120 minutes. You are asked how many of those minutes he was moving*).

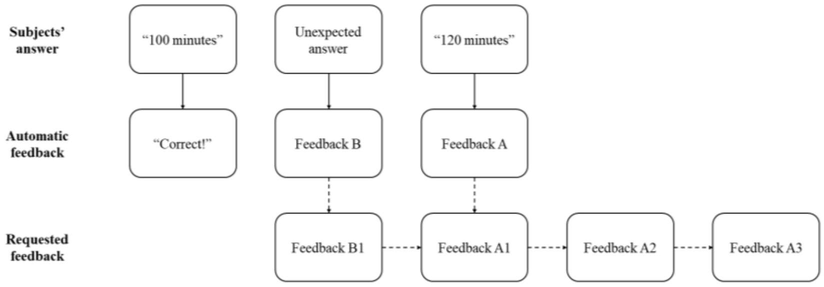


Figure 2: Offered Feedback

Data collection resulted in a STACK-generated dataset for each student attempt(s), along with cognitive interviews in which students were asked about the reasons behind their responses and reactions to feedback. Although a mixed-method approach was applied, data analysis was qualitative in nature.

Results and discussion

The analysis revealed that about one-third of students achieved the correct solution in the first attempt. The remaining ones failed on the first attempt, evidencing difficulties related to the interpretation of graphs, as reported by Graham and Sharp (1999) and Planinic et al. (2012). The first answer varied according to students' academic achievement; the medium-achievers normally made the expected error '120 min', while the low-achievers generally gave a set of unexpected responses. The cognitive interviews revealed that the latter responses stemmed from a lack of students' skills to interpret functions; for example, one of the low-achievers stated: "I was unsure about what to reply because I didn't quite understand the graph". Those who replied 120 min in the first attempt understood the task context, but they either interpreted the graph globally or did not comprehend the meaning of having scope 0. In both cases, they provided the correct answer after receiving the first feedback. During the cognitive interviews, one student stressed: "After reading the feedback, I realized that I needed to look at every part of the graph […]. The solution was not the biggest value reached in the x-axis".

Differences were also observed in how the students reacted to the feedback. Medium-achievers tended to be reluctant to ask for feedback, and instead, they reattempted the task. About half of them got the correct answer after receiving the automatic feedback (*A*), while the others required an extra hint (A1). In contrast, low-achievers were willing to demand extra information before providing a new answer, and most succeeded after receiving two extra hints. The rest did not request a third one despite not having reached the solution; they reported feeling overwhelmed and frustrated because the hints did not appear simultaneously, having to retain information from previous ones. A student reported: "Some hints were difficult to understand without thinking about the previous ones". In this case, the students were unable to solve the task, but they started to reflect on their own work, being conscious of the need to engage in learning, something already observed by Ruchniewicz and Barzel (2019). The above suggests that when designing feedback, it should be both concise and presented accumulatively, at least for low-achievers. The present results should be interpreted cautiously due to the reduced sample and the uniqueness of the task. Further research into learning graphs and e-assessments is needed to support our findings.

Acknowledgment

This work is supported by the Erasmus+ Programme of the European Union (2021-1-DE01-KA220-HED-000032031; https://www.authomath.org/). It reflects only the authors' view, and the European Commission's support does not imply endorsement of the content and is not responsible for its use.

References

Fyfe, E. R., & Rittle-Johnson, B. (2016). The benefits of computer-generated feedback for mathematics problem solving. *Journal of Experimental Child Psychology, 147*, 140–151. https://doi.org/10.1016/j.jecp.2016.03.009

Graham, T., & Sharp, J. (1999). An investigation into able students' understanding of motion graphs. *Teaching Mathematics and its Applications, 18*(3), 128–135. https://doi.org/10.1093/teamat/18.3.128

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In D. Jonassen, M. J. Spector, M. Driscoll, M. D. Merrill, J. van Merrienboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (pp. 125–143). Routledge.

Narciss, S. (2013). Designing and evaluating tutoring feedback strategies for digital learning. *Digital Education Review*, (23), 7–26. https://doi.org/10.5220/0011116400003182

Narciss, S., Prescher, C., Khalifah, L., & Körndle, H. (2022). Providing external feedback and prompting the generation of internal feedback fosters achievement, strategies and motivation in concept learning. *Learning and Instruction, 82*, 101658. https://doi.org/10.1016/j.learninstruc.2022.101658

Ortiz-Laso, Z. (2017). *Competencia matemática de los alumnos que acceden al Grado en Educación* [Master dissertation, Universidad de Cantabria].

Pinkernell, G., Gulden, L., & Kalz, M. (2020). Automated Feedback at Task Level: Error Analysis or Worked Out Examples–Which Type Is More Effective?. In B. Barzel, R. Bebernik, L. Göbel, M. Pohl, H. Ruchniewicz, F. Schacht, & D. Thurm (Eds.), *Conference on Technology in Mathematics Teaching–ICTMT 14* (pp. 221–228). Duisburg-Essen Publications.

Planinic, M., Milin-Sipus, Z., Katic, H., Susac, A., & Ivanjek, L. (2012). Comparison of student understanding of line graph slope in physics and mathematics. *International journal of Science and Mathematics Education, 10*(6), 1393–1414. https://doi.org/10.1007/s10763-012-9344-1

Ruchniewicz, H. & Barzel, B. (2019). Technology Supporting Student Self-assessment in the Field of Functions – A Design-Based Research Study. In G. Aldon & J. Trgalová (eds.), *Technology in Mathematics Teaching, Selected Papers of the 13th ICTMT Conference* (pp. 49-74). Springer.

Sangwin, C. (2013). *Computer aided assessment of mathematics*. Oxford University Press.

Development and validation of a three-dimensional framework for the classification of authentic tasks in mathematics

Marta Mikite[1], Ilze France[2] and Ģirts Burgmanis[3]

University of Latvia, Interdisciplinary Centre for Educational Innovation, Riga, Latvia;
[1]marta.mikite@lu.lv; [2] ilze.france@lu.lv;[3]girts.burgmanis@lu.lv;

Global trends in mathematics education place increasing importance on problem-solving skills in authentic contexts. In this paper, the authors propose a three-dimensional framework for the classification of authentic tasks in mathematics. These dimensions, which combine the complexity of the task but can at the same time be analyzed separately, are (1) complexity of the mathematical model, (2) context of the given problem, and (3) strategic complexity connecting the problem and the mathematical model. To validate the framework, the levels of complexity for 12 tasks are determined and students' performance on these tasks is compared. This framework is a valuable tool for designing both learning and assessment tasks.

Keywords: Authentic task, Mathematical problem, Assessment of complexity.

Introduction

Global trends in learning goals shape the Programme for International Student Assessment (PISA). PISA's latest mathematics framework highlights the importance of mathematics in today's changing world driven by new technologies. After graduation, citizens are expected to be creative and engaged, making non-routine judgments (OECD, 2023a). This means moving beyond the way mathematics is traditionally taught and prioritizing mathematics learning based on real-life examples (Kaiser & Schwarz, 2010). The latest PISA results show that Latvian students perform above average in mathematics at the lower proficiency level, but below average at the higher proficiency level (OECD, 2023b). This indicates that Latvian students need to improve their performance on problem-solving tasks in which they must think without a pre-known algorithm and in which several solutions are possible, so that they need to be more creative and evaluate their ideas. In this situation, national tests are not a driving force either. Previous analyses of Latvian national assessments show a lack of tasks with authentic contexts and show poor indicators of higher-order thinking skills. This study aims to develop and validate a multifunctional framework for designing mathematical problems with authentic contexts at different levels of complexity. A framework would help to build a common understanding of what characterizes higher-level problems in order to promote the development of higher-order thinking skills among students.

Literature review

There is no common interpretation of what is considered an authentic task or an authentic context. Some authors define it not as a property of the problem, but as a property of the connection between the problem and its solver (Kramarski, Mevarech & Arami, 2002). From this perspective, the same problem will be authentic for some students and not for others. Others stress that authenticity is determined by the fact that problems are purposeful and meaningful (Jurdak, 2006). In this study we assume that tasks in authentic contexts

> require a 'real-world' element whether in terms of meaningfulness, relevance and/or application to the personal lifeworlds of learners, as well as an element of connectedness to other subject domains and contexts beyond the textbook and school. (Tan & Nie, 2015, p. 22)

When designing mathematical tasks in authentic contexts, it is important to take into account that their content consists of multiple dimensions. Dimensions are connected in the task but can also be isolated to be analyzed separately. Pugalee and colleagues (2002) identify four dimensions: thinking and reasoning, discourse, mathematical tools, and attitudes and dispositions. Paredes and colleagues (2020) point out three main aspects that should be considered when classifying mathematics tasks: (1) the context in which the task is placed, (2) the variety of responses to the task, and (3) the level of cognitive demand activated when solving the task. Maaß (2010) has studied previously created classification versions and introduced a new, highly detailed scheme for the classification of mathematical modeling tasks. It categorizes tasks based on their characteristics and specific elements. Not all these elements affect the complexity of a task. To create assessment tasks, a framework is needed that outlines how complexity increases. It is crucial for developing an accurate assessment tool to mark the direction of intervention and improve both teaching and learning.

Methods

In this study three dimensions are distinguished which determine how complex a task is: (1) context of the given problem, (2) complexity of the mathematical model, (3) strategic complexity connecting the problem and the mathematical model (Table 1). According to the PISA 2003 Mathematics framework, each situation is more or less related to the student's world (OECD, 2003). This transfer distance forms the first dimension. The second dimension is the complexity of the mathematical model. The Structure of the Observed Learning Outcome (SOLO) taxonomy's (Biggs & Collis, 1982) unistructural, multistructural and relational levels are the basis for defining this dimension. The third dimension is about the relationship between a given situation and a mathematical model or the ability to formulate, interpret and evaluate (OECD, 2023a).

Table 1: Three-dimensional framework for a classification of authentic tasks in mathematics

	Level 1	Level 2	Level 3
Context of the given problem	A simple, straightforward, familiar situation, often in a personal context.	The situation is described using several sources of information. Although the situation is relatively familiar, it requires a deeper understanding of the context.	Complex, relatively new situation. Situation analysis or generalization is needed
Complexity of mathematical model	A simple mathematical model consisting of a single content element.	Multiple unrelated elements, an algorithm, a learned procedure.	Multiple related elements, requiring a deep understanding of mathematical concepts.
Strategic complexity connecting the problem and the mathematical model	There is a clear solution path, which may be explicitly or implicitly given in the instructions for the task. The problem allows for one correct answer.	A solution path may be chosen. There is a need to justify/explain the answer as the context allows interpretations. Assumptions need to be made.	The limitations of the context must be considered, assumptions must be made and the relevance of the mathematical model to the problem must be evaluated. The solution to the situation may differ significantly depending on the mathematical model chosen.

This study is a first validation step to test whether the complexity of the tasks created by the framework increases. The created tasks are part of the pilot study for the national numeracy monitoring in grades 6 and 7. A total of 856 participants took part in the study. The pilot study was conducted using three different item sets. The total number of items is 27, of which 6 are anchor items, identical in all tests. Items were coded based on their mathematics topic – A stands for “ratios and relationships”, G-“geometry”, L-“time and speed”, E-anchor items. The following numbers represent the task number in the students' worksheets. To ensure reliability, coefficient Cronbach's alpha was calculated for each set of results. Tasks with an authentic content were selected by experts according to the following criteria: (1) match at least the first level of the framework in each dimension, (2) fit the Rasch model. 12 tasks were selected from three item sets for the study. The Wright maps were analyzed comparing the position of different level items against the anchor items.

Results

The calculated Cronbach's alpha coefficients are 0.76; 0.67; 0.76. Considering that this is the initial pilot study, we consider these Cronbach's alpha coefficients to be acceptable to ensure reliability. In Figure 1 all the items selected for the study are framed and the determined complexity level is shown. For example, in 1/2/3 1 indicates level of the context dimension, 2 the level of complexity of the mathematical model, and 3 the level of strategic complexity.

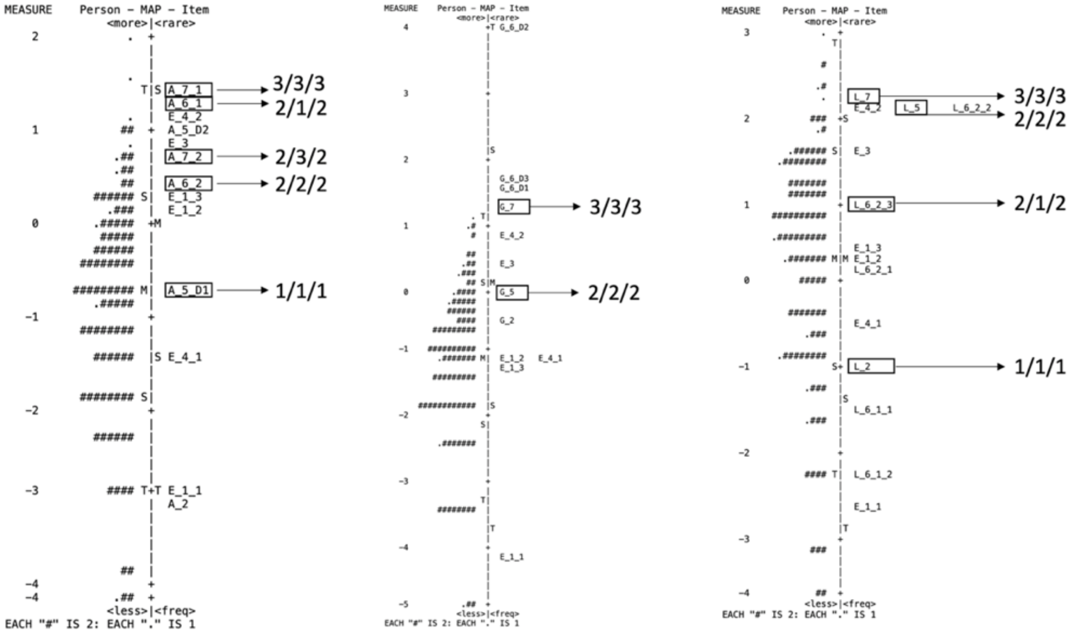


Figure 1. Item positioning in Wright maps of three item sets.

If looking at each set and each dimension separately, tasks with a higher level of complexity are positioned higher in the Wright maps, indicating that students' performance decreases with increasing levels of complexity. Item A_6_1 within the first item set does not fit the expected hierarchy in the second dimension – complexity of the mathematical model. This task requires calculating the unknown term of a proportion, which is in the curriculum at exactly the time the test is taken. Some students may have learned this skill, so it could used as a learned algorithm, but some students made up the solution in the given context.

Conclusions

The three-dimensional framework for a classification of authentic tasks in mathematics allows to purposefully increase the level of complexity. It is important to have a step-by-step approach in learning, but it is also essential in assessment to design tasks so that their complexity increases gradually to enable as many students as possible to demonstrate their best performance. It is crucial to further develop and implement the framework.

Acknowledgment

This research was supported by the project “Development of Data-Driven Solutions for School Improvement” No: lzp-2021/1-0445.

References

Biggs, J., Collis, K. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Jurdak, M. E. (2006). Contrasting perspectives and performance of high school students on problem solving in real world, situated, and school contexts. *Educational Studies in Mathematics*, *63*.

Kaiser, G., & Schwarz, B. (2010). Authentic modelling problems in mathematics education—examples and experiences. *Journal für Mathematik-Didaktik*, *31*(1), 51–76.

Kramarski, B., Mevarech, Z. R., & Arami, M. (2002). The effects of metacognitive instruction on solving mathematical authenti up the solution in the given context.c tasks. *Educational studies in mathematics*, *49*, 225–250.

Maaß, K. (2010). Classification scheme for modelling tasks. *Journal für Mathematik-Didaktik*, *2*(31), 285–311.

OECD (2003). The PISA 2003 assessment framework – mathematics, reading, science, and problem solving. Knowledge and skills. Paris: OECD.

OECD (2023a), "PISA 2022 Mathematics Framework" in PISA 2022 Assessment and Analytical Framework, OECD Publishing, Paris, https://doi.org/10.1787/7ea9ee19-en

OECD (2023b), PISA 2022 Results (Volume I): The State of Learning and Equity in Education, PISA, OECD Publishing, Paris, https://doi.org/10.1787/53f23881-en

Paredes, S., Cáceres, M. J., Diego-Mantecón, J. M., Blanco, T. F., & Chamoso, J. M. (2020). Creating realistic mathematics tasks involving authenticity, cognitive domains, and openness characteristics: A study with pre-service teachers. *Sustainability*, *12*(22), 9656.

Pugalee, D. K., Douville, P., Lock, C. R., & Wallace, J. (2002). Authentic Tasks and Mathematical Problem Solving. *Proceedings of the International Conference-The Humanistic Renaissance in Mathematics Education.*

Tan, J. P. L., & Nie, Y. (2015). The role of authentic tasks in promoting twenty-first century learning dispositions. *Authentic problem solving and learning in the 21st century: Perspectives from Singapore and beyond*, 19–39.

Fostering technology-enhanced formative assessment in Euclidean geometry proving through graded peer tutoring roles

Annamaria Miranda[1] and Loredana Saliceto[2]

[1,2] University of Salerno, Italy;[1]amiranda@unisa.it and [2] lsaliceto@unisa.it

In recent years, educators and researchers have paid increasing attention to formative assessment strategies. We explore an experimental design in which formative assessment strategies intertwine with digital technology and graded peer tutoring to overcome secondary school students' difficulties in proving Euclidean geometry statements. According to the design, each group is decomposed into three helping students, acting as guides at various levels, and one student needing to be guided to learn. Each helping student intervenes to activate a specific assessment process in a specific phase of the activity, supported or not by a digital tool, depending on the peer tutoring role to be performed. We investigate the use of digital technology and tutoring roles in supporting agents (teacher, students, peers) to develop formative assessment strategies in teaching and learning Euclidean geometry.

Keywords: Formative assessment, digital tools, roles, graded peer tutoring, proof.

Introduction and conceptual background

Formative assessment (FA) is widely regarded as one of the more effective instructional strategies employed by teachers, with a growing body of literature and academic research on the topic (Sadler, 1998; Roschelle & Pea, 2002; Irving, 2006; Wiliam & Thompson, 2007; Black & Wiliam, 2009; Swan & Burkhardt, 2014). FA refers to a wide range of methods used by teachers to conduct in-process evaluations of student understanding, learning needs, and learning progress during a lesson or course. What distinguishes an evaluation as formative is how it is used, i.e., as a method in which

> evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited. (Black & Wiliam 2009, p. 7)

Research into FA practices has particularly highlighted the role played by so-called *connected classroom technologies* (CCT), networked systems of personal computers or handheld devices specifically designed to be used in classrooms for interactive teaching and learning (Irving, 2006). CCT supports FA due to their specific features that make them effective tools for FA in accomplishing the following: (1) *monitoring students' progress, collecting the content of students' interaction over longer timespans and over multiple sets of classroom participants*; (2) *providing students with immediate private feedback, keeping them oriented on the path to deep conceptual understanding* (Irving, 2006); (3) *encouraging students to reflect and monitor their own progress* (Roschelle & Pea, 2002). Cusi et al. (2017) designed and implemented CCT-supported digital resources, the *worksheets*, to activate FA processes during classroom mathematics activities. The overall goal of FA is to collect detailed information to improve instruction and student learning while it is taking place.

We design an activity in which digital technology and metacognition support agents in activating FA processes to overcome students' difficulties in proving a statement in Euclidean geometry. A primary goal in secondary school is to have students become proficient at writing proofs in Euclidean geometry. We refer to formal proofs within the Euclidean axiomatic system as deductive arguments showing that the assumptions of a statement logically guarantee the conclusion. However, this goal is rarely met. Many causes for students' difficulties in proving seem to depend both on how to start a proof and how to bridge the gap between informal and formal reasoning (Moore, 1994; Weber, 2001). Research has also shown that using methods of informal reasoning, including visual representations, can have a positive effect on the outcome of students' proof-writing processes. Visualisation of diagrams sketching the statement of a theorem and the production of arguments, even if they are not mathematically rigorous, can lead to identifying the key idea of the proof (Raman, 2003), and the key idea begins to construct a bridge between argumentation and proof. The effectiveness of starting to prove a statement in Euclidean geometry from a visual representation is amplified if the diagram is drawn in dynamical geometry environments (DGEs). The literature regarding the use of DGEs in proof-related activity has paid attention to conjecture generation and the transition from conjecture generation to proof production (Baccaglini-Frank & Mariotti, 2010). Many studies have been conducted to investigate the affordances of DGEs, which include dragging and measuring modalities that result in the generation and testing of hypotheses by generating various diagrams (Olivero & Robutti, 2007). A statement to be proved in secondary school Euclidean geometry is frequently described with reference to a specific diagram representing a certain general class. A diagram, on the other hand, may represent only one case and thus not capture all the configurations to which the statement may refer. As a result, a diagram-based deductive proof may be valid only in that case, and different proofs may be required for different configurations. DGEs can play a significant role in this type of generalisation because their dragging function allows for easy access to multiple diagrams while maintaining the geometrical relationships imposed on the diagrams.

Our approach to achieving an efficient FA when facing a Euclidean proof intends to exploit the benefits DGEs can give students at the beginning phase of a proving process and to do this by actively engaging them in a peer tutoring setting. Moreneo & Duran (2002) describe peer tutoring (PT) as a method of cooperative learning based on the creation of pairs of students with an unbalanced relationship; that is, the tutor and the student needing help do not have equal competencies, but they share a common goal. The method can be the most intellectually rewarding experience of a student's career and serves as an effective way to improve self-esteem (Annis, 2013). Our design considers that socialisation experiences that occur during peer tutoring can benefit both the tutor and the needing student by encouraging students to learn at various levels: the interaction with more expert peers plays a crucial role in students' learning, while the tutoring role develops metacognitive competencies. This involves the development of planning, monitoring, and critiquing behaviours—all metacognitive aspects (Schoenfeld, 1992) on which FA must be focused. We report the design and implementation of a cooperative learning activity that engages secondary school students in interacting according to a peer tutoring relationship framework based on the graduation of peer tutoring by roles, with the aim to foster FA processes and help students overcome difficulties in proving in Euclidean geometry.

Theoretical framework and research questions

The theoretical framework for our design, implementation, and analysis of the activity finds its roots in the combination of the use of technology to enhance FA practices, through the three-dimensional model of FA (Cusi et al., 2017), and the role of peer tutoring in activating FA processes. Wiliam & Thompson (2007) introduce five key strategies for FA practices in school settings (WT strategies):

(a) *clarifying and sharing learning intentions and criteria for success;* (b) *engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding;* (c) *providing feedback that moves learners forward;* (d) *activating students as instructional resources for one another;* (e) *activating students as the owners of their own learning.* The three-dimensional model (Cusi et. al, 2017) considers: *the five FA key-strategies* described by Wiliam & Thompson; *the three main agents* that intervene (teachers, students, peers), and *the functionalities* through which technology can support the three agents in developing the FA strategies. The teacher, the student's peers, and the student himself or herself are the agents that activate these FA strategies. Through its three *functionalities*, the technology can assist the three agents in developing FA strategies: (1) *sending and displaying*, which fosters communication among the agents of FA processes (e.g. sending and receiving messages and files, displaying and sharing screens or documents with the whole class; (2) *processing and analysing*, which supports the processing and the analysis of the data collected during the lessons (e.g., through the sharing of the statistics of students' answers to polls or questionnaires, the feedback given directly by the technology to the students while taking tests); (3) *providing an interactive environment*, which creates environments where students can interact to work individually or in groups on tasks or explore mathematical/scientific contents (e.g. through the creation of interactive boards to be shared by the teacher and students, or through the use of specific software that provides an environment in which it is possible to explore). We design a framework in which peer tutoring (PT) is a set of pair relationships in a group graded at different tutoring levels through assigned roles. In PT, one student guides the other in conducting an assignment or learning a concept. In practice, an older student, or someone more experienced, helps a younger or inexperienced student by activating a helping process that accompanies the student in difficulty through various phases of the learning activity: understanding the assignment, exploration of the given problem, bridging to formalization, reflection, and finally evaluation. In this strand, looking at how a tutor behaves when helping a peer, we individuated the helping functions that a helper should activate to support a peer's difficulties, identifying them as specific roles. Students are engaged in group peer tutoring, where each student, except the supported student, is required to play a tutoring role at a specific level. We define a *graded peer tutoring (GPT)* within the group, that is, '*a decomposed form of peer tutoring that takes shape within the group through assigned roles at various levels, which depend on the various stages to prove the statement*'. Personifying a peer tutoring role stimulates critical reflection not only at the cognitive level, as it allows students' engagement in the mathematical problem, but also at a metacognitive level, as it fosters students' monitoring skills related to a role to play in the activity. Specifically, students playing a peer-tutoring role are forced to reflect on how one learns, learn strategies on how to learn, and, by receiving continuous feedback from the student who needs to be helped, not only monitor how she or he learns but also improve the awareness of their own learning.

We face the issue of promoting FA strategies for the development of students' proving competencies in Euclidean geometry by offering them structured tutoring opportunities that allow them to become aware of their own and others' cognitive processes, enabling them to monitor and coordinate them.

RQ: *How can digital technology and metacognitive peer tutoring roles activate formative assessment processes, helping secondary school students overcome difficulties in Euclidean geometry proofs or enhance proving competencies? Specifically, how are these factors perceived by students?*

Experimental design and definition of roles

The design of the learning activity foresees helping students face a task requiring proving a statement. Students work in groups of four, structured so that in each group there is a student needing help and three helpers interacting supported by digital technology. According to the *GPT* model we identify three levels of PT within the group, each corresponding to a specific helping phase that stimulates specific actions T1, T2, T3, and gives in-progress FA feedback (c), (d), (e): T1: *exploring and verifying;* T2: *bridging the gap between informal and formal*; and T3: *monitoring and managing the entire helping process.* Specific, prevalent, but not exclusive, corresponding FA tutoring WT strategies to be activated are: (c) -T1 *providing feedback that moves learners forward*; (d) -T2: *activating students as instructional resources for one another*; and (e) -T3 *activating students as the owners of their own learning*. This tutoring structure is well supported throughout the activity by the FA functionalities (1), (2), and (3) (Cusi et al., 2017) put in motion using digital tools. We associate actions, expected FA strategies and functionalities with specific roles. The *Jumper (J)* is the protagonist of the formative assessment activity, the receiver of help, to make a cognitive jump that reduces the knowledge gap between him and the other members of the group; the *Digital Explorer (DE),* an intermediate-level tutor, masterfully uses digital tools, helps to understand the statement and to explore the dynamical configurations of the related diagram through GeoGebra (action T1- strategy (c) - functionality (1); the *Bridging Mind (BM),* an intermediate-level tutor with a good knowledge and proof techniques, helps to bridge the gap between informal visual reasoning and formal reasoning (action T2 - strategies (c) and (d) - functionality (1) and (2)); the *Group Leader* (*GL*) a dual supervisor of the process, intervening to help other colleagues to elaborate the solution, and of the product, and has the skills to do this, high digital skills, excellent logical-deductive skills, and an aptitude for managing (action T3- strategies (c), (d), (e) - functionalities (1), (2), (3)). The assignment of the roles gives students the opportunity to find the way to stay on track (*J*), deepen their knowledge based on their awareness of what they already know and how to move on, improve their proving skills (*DE*, *BM*), learn how to learn (*GL*). The assignment is based on the students' contextual learning state ascertained by the teacher, and due to its dynamic nature, it changes as the learning states change.

Methodology

The experiment took place in a 10th-grade class of sixteen students in southwestern Italy during the school year 2017–2018, in Classroom 3.0, an environment with a flexible setting. Students were required to work in small groups, specifically four groups named $HG1$, $HG2$, $HG3$, and $HG4$, focused on collectively helping one of the components. In each group, HGi, three helpers, DEi, BMi, GLi, and the Jumper Ji were identified by the teacher according to their learning state. Specifically, the Jumper role was mainly assigned to students completely unfamiliar with the subject, having come from other classes in the previous study path. The goal was to produce proof of the following statement (Figure 1), present it and post it on a Shelf Padlet, a work to be collectively reviewed and evaluated. At the end of the activity, all students were asked to express their overall feedback on the experience, the analysis of which would have given rise to the didactic actions to be undertaken later by the teacher.

Draw a circle with diameter AB and centre C and draw the tangent lines in A and B; a third tangent at a point D of the circle intersects the other two at P and Q, respectively. Prove that PQ≅PA+QB.

Figure 1: The statement to prove

Digital tools supported all phases of the activity. Each HG had at their disposal a digital environment consisting of various tools and resources: an island station equipped with tablets connected to Internet and adjacent to traditional boards. All groups were equipped with tablets, one for each student, to explore the problem through GeoGebra; boards were used to formalise the proof; and Padlet was used to send, display, and share groups' solutions (GeoGebra and board images) and evaluations of displayed solutions presented by the jumpers; and finally, all students' individual feedback.

Data collection and Data analysis

All the data concerning the helping-learning activity has been digitally stored on shared Padlet boards. The digital environment contains the shelves where students sent and displayed on the LIM: GeoGebra files of dynamic constructions, images of the boards with the proof of the statement, evaluations of the jumpers' performances, personal feedback of the experience, and a photo gallery. Among the collected data we qualitatively analysed evaluations and feedbacks through a systematic and objective identification of some characteristics of FA processes and strategies (identified in the literature) and of the factors triggering them. Specifically, we were looking for signs of the key-FA strategies and functionalities supporting them at the evaluation phase, carried out by the groups, and at the feedback interview on students' individual immediate perceptions of the entire activity, by labelling and classifying sentences according to roles experienced. More in detail, we collected for each Jumper the groups' evaluations and, for each individual role, the impressions of the experience.

Findings and discussion

To analyse the impact that the designed components have on the activation of students' FA strategies and functionalities, we focus on the evaluations made by groups at the end and on the final interviews.

Evaluation of Jumpers by Helping Groups

In this phase, the *HG*s evaluate the *J*s's performance. Each group evaluates the other *J*s' performances according to some shared criteria: correctness and completeness of the proof, clarity of the presentation. Ji is not evaluated by the HGi, having already lived FA moments during the tutoring phases. *GL* coordinates, manages the internal discussion, *processes and analyses* (functionality (2)) the answers, and *displays* on the Padlet the evaluation expressed through a brief judgement (action T3, functionality (1)). We highlight the peer formative evaluation by looking at the FA signs in the judgments. The argument about *J1* nuances from *not very convincing, quite understandable* to *good*:

$HG2$: *J*1 *argued well* and showed that he understood the problem.

$HG3$: […] explained the proof in a way quite *understandable*.

$HG4$: […] *J*1's exposition was not very *convincing*, because of the uncertainty about the Theorem to recall, but then an intuitive hint made it clear.

Looking at $J3's$ evaluation, a lack of *self-confidence* appears, but $HG4$ argues the contrary:

$HG1$: […] has a bit of *hesitancy*, exposed the proof of the problem well.

$HG2$: […] the proof exposed by J3 was not clear, but *thanks to the help of the group*, she was finally able to prove it correctly using logical deductions.

$HG4$: […] the exposure was the most *exhaustive* and *convincing* of all.

According to *HG*s, $J4$'s presentation was the least successful; despite this, the evaluation does not highlight the failure but tends to justify it. *HG*s believe that $J4$ needs to understand that he can make the jump with another small effort, that the gap is bridgeable, and that, thanks to this experience, he has already managed to make progress and prove to himself that success is attainable. It emerges the educational value that FA strategies can have on fragile students. Students can assume more responsibility for their own learning and progresses when they are aware of their strengths and areas for improvement. Globally, evaluations *provide feedback that moves learners forward* (c)*;* and activates *students as the owners of their own learning* (e) through encouraging judgments and emphasising positive performances, with the aim of trying to remove their initial sense of inadequacy and reduce the gap between them and the tutors.

Individual feedback organized by role

All the feedbacks share an appreciation for the use of GeoGebra (functionality (3)) and the organisation in structured PT working groups, and some of them have highlighted other interesting aspects. We begin with significant excerpts from the answers of the Jumpers, the protagonists of the tutoring activity. Signs of *self-formative assessment* take shape when J_1 says he has *become more familiar with the subject*. He recognises his learning improvement, and this is feedback that makes him feel like the *owner of his own learning* (e) and *moves him forward* (c) with a new learning jump:

> *J*1: The experience in 3.0 Classroom allowed us cooperative working. It is particularly useful to self-assess and to *become more familiar with the subject*. I believe that these works, aided by digital tools, bring us into contact with the modern world.

$J2$ exhibits on a subject that he has always refused. He recognises the positive role that his comrades have played in the process as well as the validity of the experience in *moving them forward* (c),, not only for those who are in difficulty but also for *already capable* students (WT strategies (c), (d), (e)):

> *J*2: Effective and interesting experience. It has helped them expose themselves to the subject easily, thanks also to the help of friends. It has helped people with some deficiencies and strengthened some already capable knowledge!

$J3$ seems to suggest structuring the lessons in future like the one just held; both *J*3 e *J*4 emphasise the role of technologies *in providing an interactive environment* (functionality (3)) and from a collaborative perspective and the cooperation to help each other (WT strategy (d)):

> $J3$: The experience we had in the 3.0 Classroom gave us a taste of how the lessons should be carried out. You can take advantage of software like GeoGebra that shows the benefits of dynamic geometry, unlike *static geometry*, in solving a geometric problem. In addition, using online platforms increases collaboration, and joining multiple minds to create a single work will surely lead to an optimal result.
>
> *J*4: It was a genuinely nice and interesting experience. It allowed me to work better because we worked into groups so that we could compare and help each other, also thanks to the use of tablets and the GeoGebra software.

Let us now look at some feedback from helping roles *DE* and *BM*. *DE*4 emphasises the importance of the learning environment, intended not only as a physical place but also as a digital interactive place. A *DE* who does not have his own tablet appreciates moving around the classroom equipped with a device. In terms of functionality, it refers to *providing an interactive environment* (3) where students can interact to work individually or in groups on tasks or explore mathematical contents (Cusi et al., 2017), and the software *provides feedback moving students forward* (WT strategy (c)):

> $DE4$: Unlike other activities that usually take place in the classroom, it allows us to learn encouraging comparison with others and the development of more opinions.

It emerges that, although each had their own means, they all worked with digital technologies:

$BM1$: [...] we put ourselves to prove, but we also made use of new technological tools made available by the school that will certainly help us in the future [...]

With the transition from exploration with GeoGebra to axiomatic *BM3 activates an instructional resource for the Jumper* (WT strategy (d)). $BM3$, attentive to Jumper success, says:

$BM3$: The experience was incredibly positive and productive. Working in a group, each making their own contribution, made me understand that together we can quickly reach a solution. A strong point for me was the drawing conducted with GeoGebra, thanks to which the understanding of the proof was easier and more immediate. I did not find any weaknesses because every group member contributed to the work.

It is interesting to note that cooperation and technology contribute to group and individual growth:

$BM4$: [...] The possibility of cooperating and comparing each other in a constructive way contributes to collective and individual growth.

The Leader, although *engaged* in a tutoring activity (WT strategy (d)), was not bored. He no longer suggests the ordinary lessons because these are more interesting:

$GL2$: [...] interesting and engaging because it allows us to work in an unusual way from ordinary lessons and to collaborate easily thanks to the availability of the material, pushing us to work better and to attend the lessons with more interest.

GL3 fits the role of leader perfectly, as he gives feedback and moves towards learning to teach, making his colleagues move towards learning. The appreciation of novelty of the method compared to the *usual,* of the *interaction*, of the *comparison with each other* activates FA strategies (d), (e):

$GL3$: [...] A new way of teaching compared to the usual. We were able to interact more with each other and *better express our opinions* and for me it was also much easier because we worked as a group so we could help each other solve problems and explain them to each other, each with their own ideas.

A cross-cutting objective was to increase the responsibility of leaders towards the community since a knowledgeable student is often used to working alone because he or she believes that others can slow down his pace. *GL*4 captures this feedback (FA Strategy (d), (e)):

$GL4$: [...] if you are in a group with people, *you are not comfortable with* the activity is counterproductive. However, these activities can be particularly useful to help those who are having difficulty or to *reinforce* concepts that are already known.

The potential power of formative assessment for enhancing teaching and learning in mathematics education is undiscussed and strengthened using digital technology and metacognitive strategies. An attempt to solve the problem of how to help students overcome difficulties in proving Euclidean geometry can be made by designing an ad hoc activity. We design and implement an activity in which digital technology and roles are the driving forces to activate the FA process and functionalities by structuring groups according to *graded peer tutoring* in such a way to produce, at each step, instructions to move forward for the student needing help, supported by digital tools. Regarding the factors triggering FA processes, the analysis highlights that students appreciate working in structured groups, *interacting with each other by working together,* and *learning about new platforms* in a digital technology environment. Regarding the evaluation phase, *HG*s activated the following processes and strategies: *processing and analysing* the solution, *sent and displayed* on Padlet, and the presentation; *using an interactive environment* to express the evaluation; *providing* through the shared evaluation *feedback that moves learners forward;* and stimulating awareness to be *the owners of their own learning.* The first results are promising. Further research, both theoretical and practical, is needed.

References

Annis, L. F. (2013). The processes and effects of peer tutoring. *Human Learning, 10*(1), 39–47.

Baccaglini-Frank, A., & Mariotti, M. (2010). Generating conjectures in dynamic geometry: The maintaining dragging model. *International Journal of Computers for Mathematical Learning*, *15*(3), 225–253. https://doi.org/10.1007/s10758-010-9169-3

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Cusi, A., Morselli, F., & Sabena, C. (2017). Promoting formative assessment in a connected classroom environment: design and implementation of digital resources. *ZDM Mathematics Education*, Vol. *49*(5), 755–767. https://doi.org/10.1007/s11858-017-0878-0

Irving, K. I. (2006). The impact of educational technology on student achievement: Assessment of and for learning. *Science Educator*, *15*(1), pp. 13–20.

Moore, R. C. (1994). Making the transition to formal proof. *Educational Studies in Mathematics*, *27*, 249–266. https://doi.org/10.1007/BF01273731

Moreneo, C., & Duran, D. (2002). *Frameworks: Cooperative and collaborative methods.* Edebè, Barcellona, Spagna.

Olivero, F., & Robutti, O. (2007). Measuring in dynamic geometry environments as a tool for conjecturing and proving. *International Journal of Computers for Mathematical Learning*, 12(2), 135–156. https://doi.org/10.1007/s10758-007-9115-1

Raman, M. (2003). Key ideas: What are they and how can they help us understand how people view proof? *Educational Studies in Mathematics*, 52, 319–325. https://doi.org/10.1023/A:1024360204239

Roschelle, J., & Pea, R. (2002). A walk on the WILD side: How wireless handhelds may change computer-supported collaborative learning. *International Journal of Cognition and Technology*, 1(1), 145–168.Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: principles, policy & practice*, *5*(1), 77–84. https://doi.org/10.1080/0969595980050104

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. Grouws (Ed.), *Handbook for research on mathematics teaching and learning* (pp. 334–370). New York, NY: Macmillan. https://doi.org/10.1177/002205741619600202

Swan, M. & Burkhardt, H. (2014). Lesson design for formative assessment. *Educational Designer*, *2*(7), 1–24. https://nottingham-repository.worktribe.com/output/994366

Weber, K. (2001). Student difficulty in constructing proofs: The need for strategic knowledge. *Educational Studies in Mathematics*, *48*(1), 101–119. https://doi.org/10.1023/A:1015535614355

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Mahwah, NJ: Erlbaum.

CLASSIFICATION tasks as basis for assessment – Family resemblance as principle for choosing mathematical objects

Norbert Noster[1], Arnon Hershkovitz[2], Michal Tabach[2] and Hans-Stefan Siller[1]

[1]University of Würzburg, Germany; norbert.noster@uni-wuerzburg.de, hans-stefan.siller@ uni-wuerzburg.de

[2]Tel Aviv University, Israel; arnonhe@tauex.tau.ac.il, tabachm@tauex.tau.ac.il

Formative assessment is aimed at supporting learning processes. Classification tasks can be used as a basis for formative assessment, particularly for evaluating students' understanding of definitions and object properties. Here, we present six design principles for choosing mathematical objects in such tasks—across grade-levels and across the curriculum—using literature-based ideas about categories of objects, family resemblance, hierarchical structures of categories, and intuitiveness. We demonstrate how we used these principles for designing a digital classification task, and discuss further areas in which these design principles should be considered.

Keywords: Classification, Assessment, Design principles, Mathematics Education

Introduction

Learning processes in school are evaluated on a regular basis. Such evaluations that take place throughout a learning process are generally referred to formative assessment. Amongst other possible purposes it can be used to evaluate how well a learning process is going and to identify room for improvement. It can be served as aid for the teacher to plan further instructions to address issues that have been identified (Cizek, 2009). Before actions like these can take place, it is necessary to create events that serve as basis for assessment. Developing tasks that can be used for the purpose of formative assessment can be challenging, especially when trying to identify weaknesses in learners understanding of a given topic. Classification tasks have been found useful for assessing understanding (Vollrath, 1977). Utilizing these tasks, this contribution takes a conceptual and theoretical perspective aiming at providing design principles for choosing mathematical objects to be used in classification tasks for formative assessment. In this sense we use it for the purpose of testing for students' conceptions, that can serve as a tool and aid for teachers to plan following instruction. The principles focus on choosing objects for the classification task derived from ideas of classifications and relations between categories. These principles are then used to analyze an example of a digital classification task as well as findings related to that task.

Classification tasks

Classifying (also referred to as categorizing or sorting) means grouping objects that can be treated equivalent regarding a certain criterion (e.g. any object that can be used to sit on can be classified as a chair) (Rosch, 1978). It is a process that helps us structure our surroundings, for example by differentiating between edible and non-edible or living and non-living things (Richler & Palmeri, 2014; Rosch, 1978). It can be considered part or result of a learning process. Therefore, it is not surprising that it can be found in mathematics curricula all around the world (Mullis et al., 2016). Using *classification tasks*, that ask learners to classify given objects provide us with insight into their knowledge, which is why they are suitable for the purpose of assessment. Here we concur with

Vollrath (1977), who asserts the importance of classification tasks by stating that they "can make the students conscious of the characterizing properties of the concept, guide them to a definition and control their understanding of the definition" (p. 212).

Depending on the choice of objects to be classified and the choice of a classification criterion, certain features can be highlighted, and student identification of these features may be assessed. Once objects are chosen and a classification criterion is set, there are still three main options for the design of a classification tasks (Vollrath, 1978): 1) Tasks in which the objects are not classified, and the classification criterion is known, e.g., "Given is a set of quadrilaterals; classify each of them based on the existence of a line of symmetry (Has / Does Not Have)"; 2) Tasks in which the objects are classified, and the classification criterion is unknown, e.g., "Given are two groups of quadrilaterals; find the property that all the objects in Group A has and all the objects in Group B do not have"; 3) Tasks in which the objects are not classified, and the classification criterion is unknown, e.g., "Given is a set of quadrilaterals; find a property and divide the objects into two groups so that all the objects in Group A have this property and all the objects in Group B do not have it".

Different aims can be associated with different types of classification tasks, such as assessing understanding of a property, initializing development of conceptual understanding, and providing an overview of the relationship between objects and properties (Vollrath, 1977). While all three variants can be used for assessing understanding up to an extent, the first one can be used to find out to what extent a property is being identified in a set of given objects. The second one focusses more on deriving a definition from the classified objects, whereas the third variant can lead to different classifications that are not associated with the learning goal. Our focus is on the first variant as it focusses on applying a definition (of the classification criterion) rather than identifying it. Furthermore, this kind of task could be designed digitally with correct and incorrect classifications being automatically identified, so individual assessment could be automated, and immediate feedback to students could be provided (Feldt-Caesar, 2017).

For designing classification tasks, the suiting set of objects and classification criteria need to be well defined, so that they will serve the assessment purposes. This yields our first two design principles.

Design Principle 1: Define the classification criterion along with its values. This criterion should be in line with the curriculum and is strongly related to the assessment requirements. Note that classification can be done to two groups or more; for example, we can ask students to classify angles by acuteness (Yes / No) or by their type (Acute, Right, Obtuse, Straight, Reflex). Another example: For a given set of quadrilaterals, we can define a classification criterion "Has (Property of) Reflective Symmetry" (Yes / No), "Has (Property of) Rotational Symmetry" (Yes / No), or "Has This Property of Symmetry" (Reflective / Rotational / None).

Desing Principle 2: Define the set of objects to be classified. This should be in line with the curriculum and with the assessment requirements. In many cases, it may be easy to rely on pre-defined mathematical sets of objects, e.g., polygons, two-dimensional geometrical shapes, simple fractions, integers, etc. A choice should be taken at this point whether each object would belong to a single classification group. For example, if we chose the classification criterion "Has This Property of Symmetry" (Reflective / Rotational / None), and we focus on quadrilaterals, we should decide whether we want to include rectangles, which can be classified to both groups, or not.

Once classification criterion and the types of objects are chosen, we move on to choosing the specific objects that will be used in the task. For this, we present the notion of family resemblance.

Relations between categories of objects as the basis for family resemblance

Classifying objects can be complex as it can be done based on various characteristics (Pothos et al., 2011). In this context, one key term is *category*, which is defined as a group of objects that are considered to be equivalent with respect to a criterion (Rosch, 1978). Importantly, as different classification criteria could be set for the very same set of objects, categories are not given a priori. For example, in the case of classifying quadrilaterals based on the existence of a line of symmetry (Has / Does Not Have), all the rectangles would be under a single category and all the parallelograms would be under single, different category; however, if the classification criterion would be color (Red / Blue), a red rectangle and a red parallelogram would be under a single category while a blue rectangle and a blue parallelogram would be under a single, different category. Even if we limit ourselves to considering only mathematical properties, different classification criteria may yield different categories. For example, think of the set of polynomials and the following objects: x; $x + 1$; x^2, $x^2 + 1$; classifying them by the criterion "Is a Quadric Polynomial" (Yes / No) will yield the following categories: $\{x^2, x^2 + 1\}$ (Yes), $\{x, x + 1\}$ (No), while setting the classification criterion "Is a Monomial" (Yes / No) will yield different categories: $\{x, x^2\}$ (Yes), $\{x + 1, x^2 + 1\}$ (No). Objects of the same category share family resemblance (Rosch & Mervis, 1975) vis-à-vis the classification criterion. It is important to state that family resemblance refers to the extent to which an objects shares features (including irrelevant ones) with other objects of a category. From here, we derive a third design principle:

Design Principle 3: Identify categories of objects for each classification group. These categories should be in line with the curriculum and with the assessment requirements. For example, if we chose to classify based on parity of functions (Even Function / Not Even Function) and we are focused on the polynomials, possible categories for the "Even Function" group could be: monomials with even exponent, or quadric polynomials of the form $ax^2 + c$ $(with\ c \neq 0)$; categories for the "Not Even Function" could be: monomials with odd exponent, or linear polynomials.

In each of these groups, categories can be organized into a hierarchical taxonomy in which the classification criterion is inherited from a broader category to its sub-categories (Bernabeu et al., 2022; Rosch & Mervis, 1975). Constructing this hierarchy will help in identifying family resemblance, hence our fourth design principle:

Design Principle 4: Construct a hierarchical taxonomy of categories for each classification group. Think of a task for classifying functions based on parity that was presented in the previous paragraph. The two categories for the "Even Function" group could be seen as stemming from a higher-level category of polynomials with only even exponents to which another category could belong: constant polynomials.

Once the taxonomy for each classification group is set up, we can identify objects with different levels of family resemblance. The longer the path on the hierarchical structure from one category to another, the lower the family resemblance between objects of those categories. To assess understanding of a concept, it is important to include in the task objects of different characteristics, from which we derive another design principle:

Design Principle 5: For each classification group, choose objects of different levels of family resemblance.

So far, the choice of objects has been defined by their mathematical characteristics. The final step of choosing objects has to do with the students' point of view, specifically regarding their level of knowledge. For this, we regard the notion of intuition.

Intuition in mathematics education

Identifying an object as having a criterion could be done intuitively – that is, immediately, with confidence, without the need to justify this choice (Fischbein, 2002). For example, kindergarten children will identify a 3-sides polygon which has two equal sides and a third side parallel to a horizontal line as a triangle, while at the same time reject it from being a triangle if "it stands on its head" (Sinclair & Moss, 2012). Likewise, a circle would be intuitively rejected by kindergarten children from being a triangle, as it is an intuitive example for a different type of objects (Tsamir et al., 2008). Thus, an object, or a property of an object could be intuitively identified by students based on their prior experiences with these objects. In mathematics education textbooks, for example, geometrical figures are usually presented in an orientation that is parallel to horizontal and vertical lines. Hence such figures would be identified intuitively, while figures oriented differently would be less intuitive to identify (ibid). In other words, the intuitiveness of an object reflects on the level of difficulty it presents to students. For this, we yield the final design principle:

Design Principle 6: For each classification group, choose both intuitive and non-intuitive objects. Note that intuitiveness and non-intuitiveness of mathematical objects is highly sensitive to the types of objects and to the classification criterion in matter, so there cannot be general guidelines as how to design them.

Example of applying the design principles

In this section we have a look at a digital classification task that we have developed and studied (Hershkovitz et al., 2023; Noster, Hershkovitz, Siller, et al., 2022; Noster, Hershkovitz, Tabach, et al., 2022). We present it through the lens of the design principles stated above. Considering the curriculum for these grade levels in both Israel and Germany, where we studied using this task, we decided to design the task around the concept of symmetry, specifically reflective symmetry. Following Design Principle 1, we defined the classification criterion and values: Has at Least one Symmetry Line (Has / Has Not); hence this is a two-way classification task. Following Design Principle 2, we chose to focus on quadrilaterals, which are known mathematical objects for children at these ages.

In the context of quadrilaterals, there is a well-established categorization into, e.g., parallelograms, rectangles, squares, trapezoids, kites; these categories appeared in the textbooks of the populations we sampled, hence we based our design on them. Following Design Principle 3, we identified the following categories for the "Has" group: rectangles, squares, and kites; and the following groups for the "Has Not" group: Parallelograms, and Non-Isosceles Trapezoid.

Another well-established framework that was relevant to our choice of objects was the House of Quadrilaterals, which describes hierarchical relationships between different categories of quadrilaterals based on their definitions. This helped us in building taxonomies for the two classification groups. Following Design Principle 4, we constructed the following hierarchical taxonomy for the "Has" group: Squares are sub-category of Recatngles and Kites; Parallelograms and Non-Isosceles Trapezoids are sub-categories of general non-symmetrical quadrilaterals.

Now, following Design Principle 5, we chose objects to represent different levels of family resemblance. For the "Has" groups, we first chose a square and rectangle, which share high family resemblance to each other vis-à-vis reflective symmetry, and a kite, which share low family resemblance with both. For the "Has Not" group, we first chose three objects with low family resemblance between them: parallelogram, non-isosceles trapezoid, and a general non-symmetrical quadrilateral.

Finally, following Design Principle 6, we relied on a framework of intuitive and non-intuitive two-dimensional geometric objects (Tsamir et al., 2008), and added the notion that shapes with lines of symmetry that are either horizontal or vertical are more likely to be identified as symmetric than shapes with diagonal lines of symmetry (Götz & Gasteiger, 2022). Squares, rectangles, and kites are generally intuitively perceived as having a line of symmetry, while the general non-symmetry quadrilateral was built in a way that it would be intuitive to assume it had no lines of symmetry; parallelograms are non-intuitive as not having a line of symmetry, as children often mix this notion with rotational symmetry (which parallelogram do have). We added a tilted square, which is non-intuitive due to the diagonal lines of symmetry, and a rotated parallelogram, yet another non-intuitive objects, however, note that these two objects share high family resemblance with the square and parallelogram, respectively. See Figure 1.

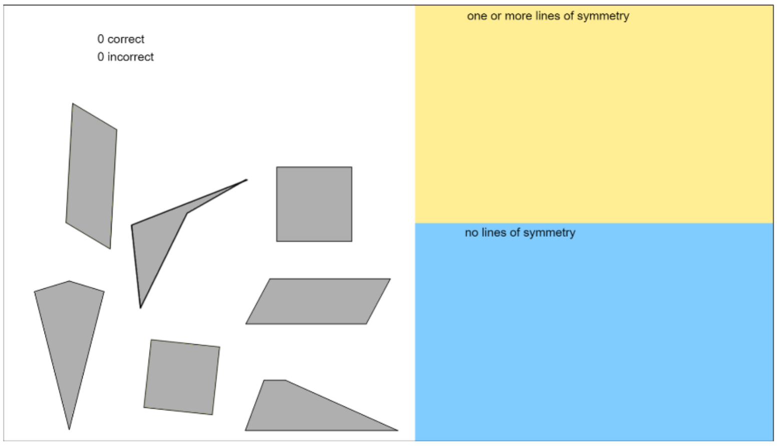


Figure 1: Examples for a digital classification task to differentiate between quadrilaterals with and without reflective symmetry

Discussion

This contribution aims at providing design principles for choosing mathematical objects to be used in classification tasks for formative assessment; these principles may be implemented across grade-levels and across topics. We focused particularly on classification tasks in which a set of objects is given along with a classification criterion, as this type of tasks provide a productive arena for assessing individual students' knowledge and can be easily assessed automatically. Based on the literature, we have identified the following design principles, that can guide the choice of the mathematical objects to be classified:

1. Define the classification criterion along with its values;
2. Define the set of objects to be classified;
3. Identify categories of objects for each classification group;
4. Construct a hierarchical taxonomy of categories for each classification group;
5. For each classification group, choose objects of different levels of family resemblance;
6. For each classification group, choose both intuitive and non-intuitive objects.

While explaining and demonstrating each principle as a "standalone", we also demonstrate the application of these principles together, to compose a classification task in a digital environment. Of course, these design principles should also be tested empirically. For example, the notion of intuitiveness of objects needs to be taken into field test, and cannot be based solely on the designers' assumptions (Noster, Hershkovitz, Tabach, et al., 2022). An iterative task design should follow, to verify the applicability of the design which resulted from applying the principles. Large data collection and its analysis could further inform the design choices.

As we are aiming at digital assessment tasks, there are also other aspects to be considered, to which we have not referred. These include for example the issue of feedback use. A digital environment could provide feedback of different types: simple vs. elaborated; immediate vs. delayed; or feedback on correctness vs. on strategies being implemented (Attali & van der Kleij, 2017; Shute, 2008; Tärning, 2018). To the best of our knowledge, there are no conclusive guidelines in the literature as for the most effective combination of these options for the purpose of serving assessment. This is an avenue for further research we are planning to pursue.

Another issue to be considered and tested relates to the layout of the objects on the screen. This may involve aspects like arrangement of the classification areas, e.g., unclassified objects are placed in the middle, in between the classification areas vs. to the left or right of the classification areas; arrangement of the objects, e.g., on a grid vs. randomly located; or issues related to size and color. We consider it an empirical question that needs to be tested based on large data collection, preferably in a set of randomized controlled studies, which again could lead to an iterative process of refining the design. To this we shall add more traditional design issues related to user interface, which are crucial in digital learning environments (Park & Song, 2015; Sagrario & Simbulan, 2007).

While the considerations presented here provide principles for choosing objects for classification tasks, it should be stated that this is only the first step in task design, as depending on the concrete task the effects may vary. Task design should be iterative (Liljedahl et al., 2007), but our principles serve as a foundation for a first predictive analysis, which has proven to be useful in the reflection of the data.

Acknowledgment

We thank the Bavarian Research Alliance and their Funding Programme for the Initiation of International Projects (BayIntAn) as well as their Scholarship Program for Bavarian and Israeli Early Career Researchers for supporting this project.

References

Attali, Y., & van der Kleij, F. (2017). Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. *Computers & Education*, *110*, 154–169. https://doi.org/10.1016/J.COMPEDU.2017.03.012

Bernabeu, M., Moreno, M., & Llinares, S. (2022). Preservice primary teachers' curricular reasoning when anticipating primary students' answers to geometrical figure classification tasks. *Proceedings of the Twelfth Congress of the European Society for Research in Mathematics Education*. https://hal.science/hal-03751484

Cizek, G. J. (2009). An introduction to formate assessment: History, characteristics, and challenges. In H. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 3–17). Routledge. https://doi.org/10.4324/9780203874851-6

Feldt-Caesar, N. (2017). Konzeptualisierung und Diagnose von mathematischem Grundwissen und Grundkönnen. *Konzeptualisierung Und Diagnose von Mathematischem Grundwissen Und Grundkönnen*. https://doi.org/10.1007/978-3-658-17373-9

Fischbein, E. (2002). *Intuition in Science and Mathematics: An educational approach*. Kluwer Academic Publishers. https://doi.org/10.1007/0-306-47237-6

Götz, D., & Gasteiger, H. (2022). Reflecting geometrical shapes: Approaches of primary students to reflection tasks and relations to typical error patterns. *Educational Studies in Mathematics*, *111*(1), 47–71. https://doi.org/10.1007/s10649-022-10145-5

Hershkovitz, A., Tabach, M., Noster, N., & Siller, H.-S. (2023). Student behavior while engaged with feedback-enhanced digital sorting tasks. In M. Ayalon, B. Koichu, R. Leikin, L. Rubel, & M. Tabach (Eds.), *Proceedings of the 46th Conference of the International Group for the Psychology of Mathematics Education (Volume 3)* (pp. 51–28). PME.

Liljedahl, P., Chernoff, E., & Zazkis, R. (2007). Interweaving mathematics and pedagogy in task design: A tale of one task. *Journal of Mathematics Teacher Education*, *10*(4–6), 239–249. https://doi.org/10.1007/s10857-007-9047-7

Mullis, I. V. S., Martin, M. O., Goh, S., & Cotter, K. (Eds.) (2016). TIMSS 2015 Encyclopedia: Education Policy and Curriculum in Mathematics and Science.

Noster, N., Hershkovitz, A., Siller, Hans.-S., & Tabach, M. (2022). Students' strategies for identifying reflective symmetry of extra-mathematical shapes in a digital environment. *ERME Topic Conference on Mathematics Education in the Digital Age*.

Noster, N., Hershkovitz, A., Tabach, M., & Siller, H.-S. (2022). Learners' strategies in interactive sorting tasks. In I. Hilliger, P. J. Muñoz-Merino, T. De Laet, & A. F. T. Ortega-Arranz (Eds.), *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption. EC-TEL 2022. Lecture Notes in Computer Science, vol 13450* (pp. 285–298). Springer.

Park, H., & Song, H. D. (2015). Make e-learning effortless! Impact of a redesigned user interface on usability through the application of an affordance design approach. *Educational Technology & Society*, *18*(3), 185–196.

Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Edwards, D. J., Hines, P., & McDonnell, J. V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition*, *121*(1), 83–100. https://doi.org/10.1016/j.cognition.2011.06.002

Richler, J. J., & Palmeri, T. J. (2014). Visual category learning. *WIREs Cognitive Science*, *5*(1), 75–94. https://doi.org/10.1002/wcs.1268

Rosch, E. (1978). Principles of Categorization. In E. Rosch & B. B. Lloyw (Eds.), *Cognition and Categorization* (pp. 27–48). Lawrence Erlbaum.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Sagrario, M., & Simbulan, R. (2007). Learning objects' user interface. In A. Koohang & K. Harman (Eds.), *Learning objects: Theory, praxis, issues, and trends* (pp. 259–336). Informing Science Press.

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Sinclair, N., & Moss, J. (2012). The more it changes, the more it becomes the same: The development of the routine of shape identification in dynamic geometry environment. *International Journal of Educational Research*, *51–52*, 28–44. https://doi.org/10.1016/j.ijer.2011.12.009

Tärning, B. (2018). Review of feedback in digital applications - Does the feedback they provide support learning? *Journal of Information Technology Education: Research*, *17*, 247–283. https://doi.org/10.28945/4104

Tsamir, P., Tirosh, D., & Levenson, E. (2008). Intuitive nonexamples: The case of triangles. *Educational Studies in Mathematics*, *69*(2), 81–95. https://doi.org/10.1007/s10649-008-9133-5

Vollrath, H. J. (1977). The understanding of similarity and shape in classifying tasks. *Educational Studies in Mathematics 1977 8:2*, *8*(2), 211–224. https://doi.org/10.1007/BF00241026

Vollrath, H. J. (1978). Klassifikation nach Ähnlichkeit. *Der Mathematikunterricht* , *24*(2), 105–115.

Effects of a teacher professional development program in formative assessment on teachers' conceptions of feedback and assessment and their self-reported feedback practices

Torulf Palm[1], Gavin Brown[2,3], and Björn Palmberg[4]

[1]Umeå University, Umeå Mathematics Education Research Centre (UMERC), Department of Science and Mathematics Education, Sweden; torulf.palm@umu.se

[2]Umeå University, Department of Applied Educational Science, Sweden; gavin.brown@umu.se

[3]The University of Auckland, Faculty of Education and Social Work, New Zealand; gt.brown@auckland.ac.nz

[4]Umeå University, Umeå Mathematics Education Research Centre (UMERC), Department of Science and Mathematics Education, Sweden; bjorn palmberg@umu.se

Formative assessment (FA) has been shown to have the power to improve student achievement. Therefore, many professional development (PD) initiatives have been carried out to support teachers to develop their FA practice. However, accomplishing such practices have been proven difficult. Among factors that are important for outcomes of PDs are teachers' beliefs and conceptions. This study examines the effects of a PD in FA on teachers' conceptions of assessment and feedback, and self-reported feedback practices. These variables were measured through a survey in the beginning and at the end of the PD. Differences between the intervention and a control group were examined at both time points using factor analytic methods and with t-tests on change scores. For the intervention group, significant positive differences were found in both the means of important conceptions and in the strength of relationships, while this was not the case for the control group.

Keywords: Beliefs, conceptions, feedback, formative assessment, professional development.

Introduction

Formative assessment (FA) is a classroom practice in which teachers and/or students elicit evidence of students' learning needs through assessment and then adapt teaching and/or learning accordingly. It has been shown to have the potential to improve student achievement (e.g., Baird et al., 2014). However, although some professional development programs (PDs) have succeeded in helping teachers accomplish FA practices that improve student learning (e.g., Andersson & Palm, 2017), they have often been unsuccessful in accomplishing substantial improvements in teachers' FA (e.g., Randel et al., 2016). Teachers' conceptions of, or beliefs about, assessment and feedback are among factors that affect implementation of FA components (Brown et al., submitted). However, studies examining effects of PDs on teacher conceptions of assessment and feedback are rare, but e.g., Deneen and Brown (2016) did not find effects from a course on assessment literacy on teachers' conceptions of assessment. This paper focuses on the effects of a PD program in FA on mathematics teachers' conceptions of assessment and feedback, and self-reported feedback practices.

Methods

Design

The PD ran as an experimental intervention with control group. A pre- and post-experiment survey was conducted in a northern Swedish city with a large control group and a small experimental group. Differences between groups were examined at both time points using factor analytic methods.

Participants

A total of 461 teachers working between school years 1 and 9 responded to the survey. Among them, 257 teachers responded to the survey at both times. They were matched between time 1 (2021) and time 2 (2023) so the variance over time could be properly evaluated. Expectation maximation was used to impute the small amount of missing data.

Professional development intervention

The PD was designed based on reviews on characteristics important for PD outcomes (e.g., Heitink et al., 2016) and on our own previously arranged PDs (e.g., Andersson & Palm, 2017). It was organized by a research team led by the first author. The researchers and the teachers met once a month during 3-6 hours for three years. The teachers also met by themselves once a month. The meetings included lectures about FA and concrete activities for its implementation, as well as group discussions and analysis of the content and suggested activities. General FA strategies were concretized for mathematics (e.g., what effective questioning would look like in mathematics). Time was also put aside for the teachers to plan for implementing FA activities in their classrooms. In the following meeting the teachers evaluated the try-outs, shared experiences of success and discussed how they could overcome obstacles and develop the use of a particular activity. The researchers supported these discussions and intervened with suggestions when deemed useful. The teachers were supported in their self-regulated learning of FA by being provided an evaluation tool, and time to use it, for evaluating and setting goals for their practices. Generally, the programme possessed a formative, process-oriented character and provided support for the teachers to influence the program.

Instruments

The Swedish Teachers Conceptions of Assessment inventory (TCoA) measures three major constructs. For the purposes of this study, only the conception that assessment improves teaching and learning (Assessment Improves) was selected as it was most sensitive to the impact of the formative assessment PD program and is likely to facilitate implementation of FA. This Assessment Improves factor has 4 1st-order factors (i.e., assessment helps teachers improve teaching, assessment helps students improve learning, assessment is reliable, and assessment is diagnostic).

The Swedish Teachers Conceptions of Feedback inventory (TcoF) (Brown et al., 2023) consists of six feedback conceptions and a self-reported Formative Feedback Practices factor. The Formative Feedback Practices factor is predicted by the two feedback conceptions Feedback Improves Performance and Students Ignore Feedback. Of the seven factors in the TcoF, only four were retained in this study (i.e., Students Ignore Feedback, Feedback Improves Performance, Feedback Involves Students in Peer and Self-feedback, and Formative Feedback Practices) as they were most likely to be sensitive to the professional development and are likely to facilitate implementation of FA.

Data analysis

The model we used had assessment conceptions predicting feedback conceptions and practices on the assumption that feedback generally occurs after assessment events (Hattie & Timperley, 2007). To account for the repeated measures design, a cross-lagged, bivariate path model with autoregressive paths (Curran & Bollen, 2001) was tested. Within each time point, the assessment conceptions factor (Assessment improves) with four dependent scores was regressed onto a general feedback factor. This general feedback factor had three dependent scores from the retained feedback conceptions factors, and was regressed onto the Formative Feedback Practices factor. A path from Student Involvement in Feedback factor to Formative Feedback Practices was added. Autoregressive paths from each variable at Time 1 were added to the matching variable in Time 2. No cross-lag paths from Assessment or Feedback factors at Time 1 to Time 2 could be identified. Hence, the model could be described as a structural path model within time with auto-regressive paths across time. To compare the model between the two groups, nested invariance testing was conducted (Brown et al., 2017). Structural equation modeling (SEM) and invariance testing were conducted with AMOS v29.0.0 (IBM, 2022).

Results

Prior to multi-group analysis, the assessment to feedback model with autoregression was found to have acceptable to good fit for the whole group. Also, the input model for the two-group analysis had acceptable fit. Invariance testing showed that measurement weights were not equivalent between groups. Hence, the two groups differed at the unconstrained level, indicating that they were drawn from two separate populations.

The intervention group differed from the control group in significant and substantial ways that were most notable after the intervention itself. Based on a t-test of differences on the change score from pre- to post-intervention time point, the intervention group gained substantially (Cohen's $d \geq .50$) for Assessment Helps Students Improve; Assessment is Reliable, Feedback Improvement, and Formative Feedback Practices variables. These differences in changes were due to that the intervention group's conceptions of assessment and feedback as well as their feedback practices moved substantially in favour of formative assessment, while the means for the control group fundamentally remained constant. Equally notable, there was a positive shift in the strength of the relationships from the assessment and feedback latent factors to their respective items only in the intervention group. Throughout, the control group, as would be expected without any focused professional development, did not change in means or model path values.

Discussion

In contrast to the lack of effects found from the assessment course studied by Deneen and Brown (2016), the PD in this study improved teachers' conceptions of assessment and feedback, and their formative feedback practices. Which PD features that were decisive for the effects cannot be determined from the study. However, giving the teachers both time and support for planning implementation in their classes together with support to overcome difficulties may have played a role. Also, the rather substantive length of the PD may have been a factor since belief change often takes time and occurs gradually. These are features found to be important for positive outcomes of PD in FA (Heitink et al., 2016). Finally, the formative character of, and the support for teachers to influence, the PD together with support for taking individual and collective responsibility for their own learning through self-regulated learning processes may have contributed to the positive outcomes. Beliefs are commonly measured through questionnaires, but a limitation of this study is that the teachers' practices are self-reported and not observed. Future studies that include observations of teacher practices and a focus on which PD features that are decisive for outcomes would be valuable.

References

Andersson, C., & Palm, T. (2017). The impact of formative assessment on student achievement: A study of the effects of changes to classroom practice after a comprehensive professional development programme. *Learning and Instruction, 49*, 92–102. https://doi.org/10.1016/j.learninstruc.2016.12.006

Baird, J., Hopfenbeck, T., Newton, P., Stobart, G., & Steen-Utheim, A. (2014). *State of the field review: Assessment and learning*. Oslo: Report for the Norwegian Knowledge Centre for Education, case number 13/4697. Retrieved from http://forskningsradet.no

Brown, G. T. L., Andersson, C., Winberg, M., & Palm, T. (2023). Predicting formative feedback practices: Improving learning and minimising a tendency to ignore feedback. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1241998

Brown, G. T. L., Harris, L. R., O'Quin, C., & Lane, K. E. (2017). Using multi-group confirmatory factor analysis to evaluate cross-cultural research: identifying and understanding non-invariance. *International Journal of Research & Method in Education*, *40*(1), 66-90. https://doi.org/10.1080/1743727X.2015.1070823

Curran, P. J., & Bollen, K. A. (2001). The best of both worlds: Combining autoregressive and latent curve models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 107–135). American Psychological Association. https://doi.org/10.1037/10409-004

Deneen, C. C., & Brown, G. T. L. (2016). The impact of conceptions of assessment on assessment literacy in a teacher education program. *Cogent Education*, *3*, 1225380. https://doi.org/10.1080/2331186X.2016.1225380

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81e112. http://dx.doi.org/10.3102/003465430298487 .

Heitink, M. C., Van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K., & Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educational research review*, *17*, 50-62.

Randel, B., Apthorp, H., Beesley, A., Clark, T., & Wang, X. (2016). Impacts of professional development in classroom assessment on teacher and student outcomes, The Journal of Educational Research, 109:5, 491-502, https://doi.org/10.1080/00220671.2014.992581

Analytical readings' method as a tool to assess the word problem-solving process involving rational numbers

Sofia Peregrina Fuentes[1], Carlos Valenzuela García[2], María Teresa Sanz[3] and Emilia López-Iñesta[4]

[1]University of Guadalajara, Mexico; sofia.peregrina5508@alumnos.udg.mx

[2]University of Guadalajara, Mexico; carlos.valenzuela@academicos.udg.mx

[3]University of Valencia, Spain; m.teresa.sanz@uv.es

[4]University of Valencia, Spain; emilia.lopez@uv.es

Word problem solving in mathematics is essential, and the processes that students follow to solve them may vary. Consequently, teachers face the challenge of interpreting these strategies, while students sometimes have difficulties explaining their processes. The analytical readings' method is presented as a resource for the assessment of the word problem-solving processes, in addition to helping students develop skills in this area. In this context, the purpose of this study is to document reflection, especially among undergraduate mathematics students with an interest in teaching, on analytical readings as a tool to assess problem solving. Reflections highlight that the analytical readings' method allow assesses students' comprehension of the problem, their performed operations, the reasoning behind them, providing a comprehensive view of the thinking process.

Keywords: Word problems, analytical readings, word problem-solving process, assessment, rational numbers.

Introduction

Word problem-solving is considered a basic competence that promotes the development of arithmetic and algebraic thinking (Siegler et al., 2013). Rodríguez (2012) indicated that when solving a word problem, the student is allowed to "explore, experiment, analyse their progress, change course, reflect on what they have done, notice how they are thinking and approaching the task, etc." (p. 154).

Currently, word problems are an essential part of the mathematics education of every student around the world, and consequently, teachers must have resources to assess the solving process of these word problems. Therefore, our main objective is to promote the use of the analytical readings' method as a tool to assess the process of solving word problems with rational numbers, for this purpose the reflections of undergraduate mathematics students interested in teaching were documented.

It is clarified that it is not intended to promote the method as the best option or to say that it is the most efficient way of evaluation, but rather it is presented as another tool to evaluate, with advantages and disadvantages, which are documented in this study.

Theoretical framework

A word problem is a verbal description of a situation in which a question is posed and the answer to which can be found by applying mathematical procedures to the numerical data provided in it (Verschaffel et al., 2020). The word problem-solving process is understood as the student's mental activity from the moment he encounters a problem that must be solved until the task is finished;

considering ideas from Polya (1957), Puig and Cerdán (1988) proposed six phases in the resolution process: reading, comprehension, translation, calculation, solution, and verification.

The reading and comprehension phases were defined separately to emphasize the attention that should be placed on reading the problem at the beginning of word problem-solving instruction. However, reading and comprehension are not independent of each other, since they are aspects of the same operation that has the purpose of understanding the word problem.

The translation phase consists of the identification of the variables involved in the word problem, both known and unknown, and the relationship between them. In this way, three important aspects must be considered: what data is going to be handled, what operations or procedures will be carried out, and in what order.

The calculation phase refers to the execution of operations and algorithms. In this phase, the student's translation skills no longer intervene, but rather their algorithmic skills. It is important to note that Puig and Cerdán (1988) consider that "translation and algorithmic skills are usually independent of each other" (p. 14).

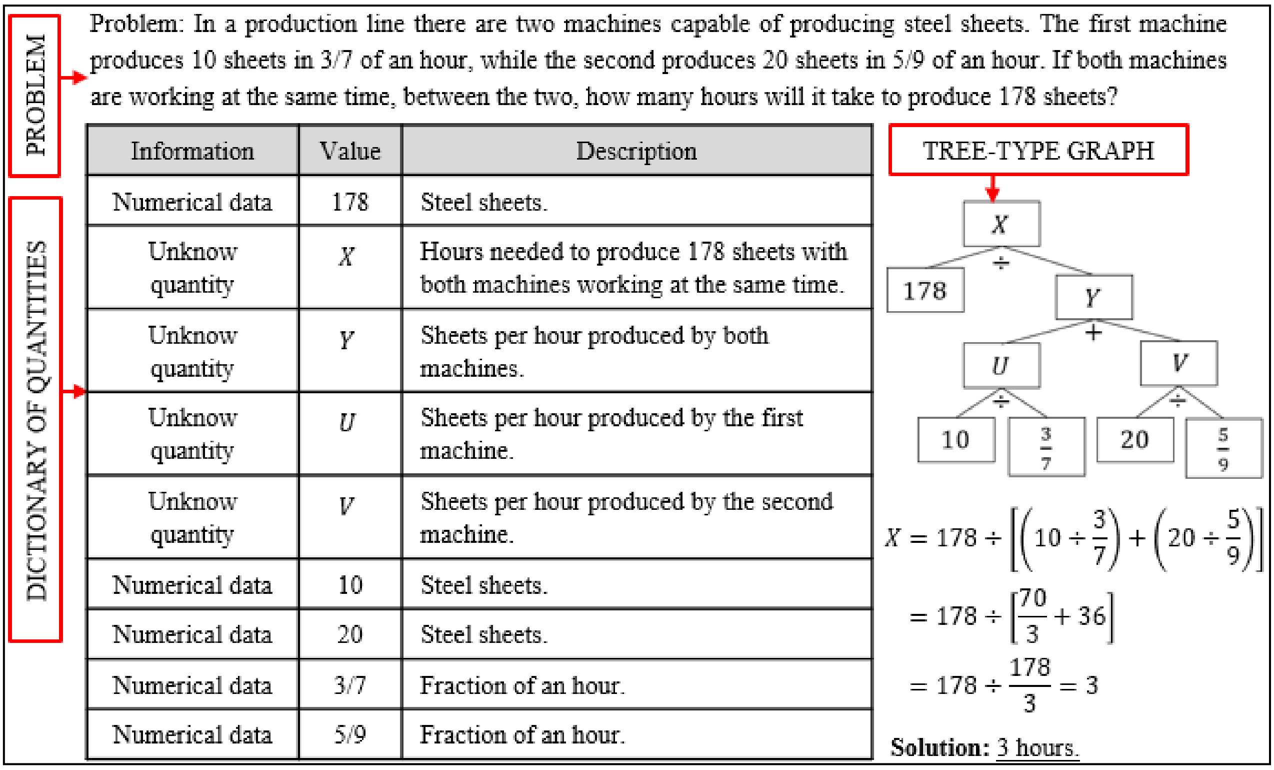
Finally, the solution phase consists of interpreting the numerical result obtained in terms of what is asked in the word problem, and in the verification phase we proceed to verify that said solution is adequate. The verification can range from something informal, such as seeing if the result makes sense (i.e. that there are no negative distances), to something more formal, such as substituting a value in an equation or solving the problem using a different procedure and compare the results.

One of variables associated with the word problem-solving process is the content variable. This is related to the mathematical meaning of the problem. The content variable in this study is related to rational number as operator. These numbers are associated with diverse uses. Based on the phenomenological analysis of fractions proposed by Freudenthal (1983), and the interpretation of Valenzuela (2018), five uses of fractions are distinguished at an abstract level: as fracturer, comparer, measurer, operator, and number. As said before, this work focuses on the operator aspect, that is, when the fraction acts on a quantity by expanding or reducing it. The operator aspect of the fraction is related to the phenomena of reproducing, reducing, enlarging, shrinking, expanding, contracting, etc. This operator aspect of the fractions can be extended to the diverse ways to express a rational number, such as decimal notation, percentage, and ratios.

Analytical readings' method

Taking as reference the analysis-synthesis method and the Cartesian method (Puig & Cerdán, 2014), the analytical readings' method to solve word problems is proposed, which is used indifferently for algebraic or arithmetic resolution processes. The method consists of the following five steps: reading the problem, forming a dictionary of quantities, building a tree graph, calculation, and solution-verification (see Figure 1).

In the first step, a first reading of the complete problem is done. Then it is read sentence by sentence, identifying the data presented and the verbs associated with them. Also, the main unknown quantity must be identified, that is, what needs to be calculated.



Problem: In a production line there are two machines capable of producing steel sheets. The first machine produces 10 sheets in 3/7 of an hour, while the second produces 20 sheets in 5/9 of an hour. If both machines are working at the same time, between the two, how many hours will it take to produce 178 sheets?

Information	Value	Description
Numerical data	178	Steel sheets.
Unknow quantity	X	Hours needed to produce 178 sheets with both machines working at the same time.
Unknow quantity	Y	Sheets per hour produced by both machines.
Unknow quantity	U	Sheets per hour produced by the first machine.
Unknow quantity	V	Sheets per hour produced by the second machine.
Numerical data	10	Steel sheets.
Numerical data	20	Steel sheets.
Numerical data	3/7	Fraction of an hour.
Numerical data	5/9	Fraction of an hour.

$$X = 178 \div \left[\left(10 \div \frac{3}{7}\right) + \left(20 \div \frac{5}{9}\right)\right]$$

$$= 178 \div \left[\frac{70}{3} + 36\right]$$

$$= 178 \div \frac{178}{3} = 3$$

Solution: 3 hours.

Figure 1: Example of the Analytical readings' method

To form the dictionary of quantities, the known and unknown variables that appear in the word problem are identified. This is organized in a table with three columns. The first column indicates whether the information is a given numerical data or an unknown value that must be calculated. In the second column, the numerical value of the data or what the unknown variable will be called is indicated. In the third column, a brief description of the variable is given.

The third step is to build the tree-type graph (Figure 2) following the rules of analysis-synthesis so that at the end all the numerical data in the problem and the main unknown quantity are obtained.

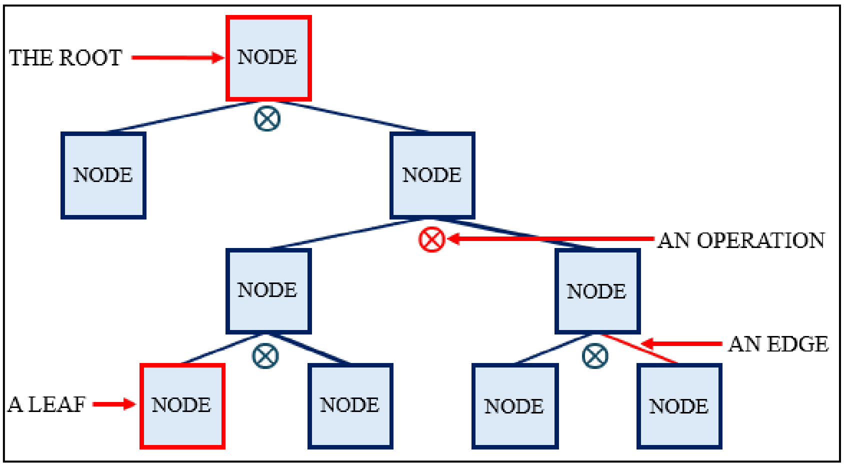


Figure 2: Tree-type graph

A tree-type graph has certain unique decomposition properties that allow obtaining a visual reproduction of the strategy followed in the resolution process (Roy & Roth, 2015). A tree is a data structure consisting of nodes connected by edges, with the unique property of not containing cycles.

For the analytical readings' method, we specifically use a rooted binary tree, in which there is a topmost node, commonly known as the root of the tree, and each node is linked to zero or two successor nodes. The nodes without successor nodes are called leaves.

In the fourth step of the method, the calculations that appear in the graph will be conducted, in the order that it shows. Finally, the fifth step is to write the result in terms of what is asked and verify it. In the Figure 1, an example of the steps followed in the analytical readings' method to solve a problem in shown. It should be noted that during the implementation of the analytical readings' method, the transit between its steps should not follow a strict linear order, it is always possible to go back and forth between the different stages.

Methodology

This qualitative research documents how mathematics students conceive the analytical readings' method as a tool to assess the word problem-solving process. In the study 13 students of the bachelor's degree in mathematics at the University of Guadalajara were considered. This is a group with affinity for teaching. The intervention consisted of three sessions that were part of an optional subject offered to the students. The worksheets, surveys, and essays were considered for the analysis. These were complemented with field notes, audio, and video recordings.

In the first session, the analytical readings' method was studied and explored through a document that simulated the solution of a word problem solved by a student -Paula-, reflecting in the process the dictionary of quantities, the tree-type graph, and its solution. The problem in question is the one shown in Figure 1. For this exploration, teams of 3 or 4 members were formed. Teams are instructed to discuss what the student meant by what she illustrated, what her procedure was, and whether her answer is correct. Afterwards, their reflections were presented to the class. Next, the phases in the word problem-solving process proposed by Puig and Cerdán (1988) were explained. The steps that make up the analytical readings' method were delineated, while inviting the students to share their observations, contributions, and doubts. In this session, two word problems were given to the students to solve and practice. For each problem, a student goes to the board to solve it using the method learned, with comments and suggestions from the other students and the teacher's guidance.

In the second session, the students began by solving a problem in teams through the analytical readings' method as a review, comparing and discussing the results as a group. Subsequently, the different aspects of rational numbers and diverse forms to express them were exposed and discussed. Four problems were answered in teams using the diverse forms to express the rational numbers studied using the analytical readings' method (word problems involving fractions, ratios, percentages, and decimal notation). The dictionary of quantities, tree-type graph and solution of each team were compared and discussed.

In the third session, the relationship between the analytical readings' method and the word problem-solving process was discussed in a debate moderated by the professor. And finally, an anonym survey was carried out to identify students' opinions on the use of the method.

As a final assignment, the students were asked to write an essay answering the following questions: 1) How is the solver's thought process reflected when solving a word problem in each of the steps of the analytical readings' method? 2) How is each phase of the problem-solving process related to the steps of the analytical readings' method? 3) What is the usefulness of the analytical readings' method as an assessment tool for teachers?

Results

The results are divided into three sections: the relationship between the steps of the analytical readings' method and the phases of the word problem-solving process identified by the students, their opinions on the advantages and disadvantages of the method as an assessment tool, and their views on how this method adapts for word problems that involve each form to express the rationale numbers seen during the intervention.

Relationship between the steps of the analytical readings' method and the phases of the word problem-solving process

As described in the methodology, during the third session, the relationship between the analytical readings' method and the problem-solving process was discussed by the students. The relationships found by the students are summarized in Table 1.

Table 1: Relationship between the steps of the method and the phases of the problem-solving process

Steps of the analytical readings' method	Phases in the word problem-solving process that are related
Reading the problem	The reading and comprehension phases are related with this step of the analytical readings' method since the student is instructed to identify the question posed and the main unknown quantity, that is, what you should be calculated.
Forming a dictionary of quantities	The comprehension phase is also related with the step of forming a dictionary of quantities, because it helps to define the starting point (given numerical data) and know how to classify the information that you consider should be used to reach the solution. Also, a fundamental part of the translation phase is the identification of the variables involved in the word problem, which are reflected in the dictionary of quantities in an orderly manner.
Building a tree graph	Other aspects of the translation phase are what operations or procedures will be carried out and in what order? these are reflected in the tree graph.
Calculation	The step of calculation in the analytical readings' method, where the calculations that appear in the tree graph are conducted in the in the order that it shows, is equivalent to the phase of calculation in the problem-solving process, when operations and algorithms are executed.
Solution-verification	The solution-verification step of the method is analogous to the solution and verification phases of the word problem-solving process.

Advantages and disadvantages of the method as an assessment tool

During the second session, when the teams presented and compared their dictionary of quantities, tree-type graph and solution of four different problems, each one using the diverse forms to express the rational numbers, a group dynamic of assessment was carried out. Each team tried to describe the thought processes of their partners by reading the method. Subsequently, each student gave their opinion on the advantages and disadvantages of the analytical readings' method as an assess tool in their essays and anonymously in the survey. These opinions are summarized in the Table 2.

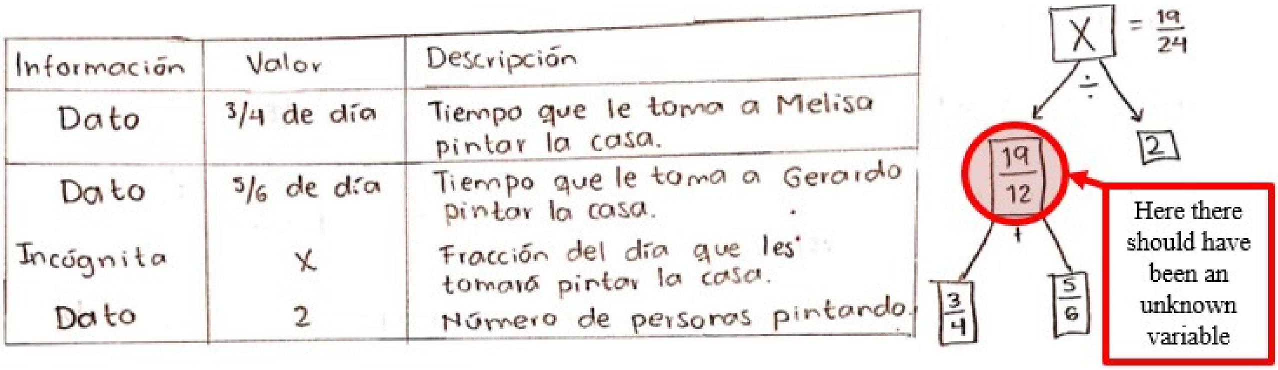
Table 2: Advantages and disadvantages of the analytical readings' method as an assessment tool

Advantages	Disadvantages
This method works as a powerful assessment tool, through which it is not only possible to assess mathematical knowledge but also to assess reasoning, logic and even reading comprehension skills. This tool can provide extra information about the students' thinking process. As teachers, the steps of the analytical readings' method help us identify the student's points of deficiency. The incorrect construction of the dictionary of quantities suggests problems with the reading and/or understanding of the text. The erroneous layout of the graph indicates problems with the translation of the text into mathematical processes and the incorrect calculation of the data reveals deficiencies in the student's arithmetic knowledge. In addition, the method can be very useful for teachers when grading, as it encourages students to express their procedure in an orderly manner.	The analytical readings' method can be laborious due to the details to be specified in each step; this mainly has two disadvantages: By virtue of its slow nature, the method is restricted to limited uses in a classroom environment where deadlines must be met. Before it can be used as an assessment tool, students must be taught to work with this method. When the method is introduced for the first time, the solution process can be confusing at first, especially the tree-type graphs, since it is a new way of representing operations. Therefore, extra sessions would be needed for the group to become familiar with the method.

An example, when discussing the procedure made by a team to solve the next problem:

> Problem 1. Melisa and Gerardo are going to paint the house where they both live. Melisa can paint the house in 3/4 of a day, while Gerardo is able to paint the house in 5/6 of a day. If they work together, what fraction of a day will it take them to paint the house?

The students highlighted the importance of seeing the variables in the dictionary of quantities. They argued that if the variable written in the tree graph (Figure 3) had been written in the dictionary, it would have allowed them to know how their classmates were interpreting the sum of the fractions of days that Gerardo and Melisa took to paint the house individually, why they carried out this operation, and therefore better understand what the error was in understanding the problem that led to the wrong answer. During the discussion, it was also mentioned that the tree graph allowed them to see how the fractions 3/4 and 5/6 were related, which led their classmates to an error.



Información	Valor	Descripción
Dato	3/4 de día	Tiempo que le toma a Melisa pintar la casa.
Dato	5/6 de día	Tiempo que le toma a Gerardo pintar la casa.
Incógnita	X	Fracción del día que les tomará pintar la casa.
Dato	2	Número de personas pintando.

X = 19/24
÷
19/12
2
+
3/4
5/6

Here there should have been an unknown variable

Figure 3: Procedure made by a team while solving problem 1

The use of the analytical readings' method for solving word problems that involve diverse forms to express the rational numbers.

As indicated previously, during the intervention the students worked with four different word problems, which involved fractions, ratios, percentages, and decimal notation. In the survey, participants were asked to indicate how well they think the method is adapted to solve problems involving each of these four ways of expressing rational numbers, the options offered were: 1) *It does not fit well* -the explanation of my procedure feels forced when I use this method-. 2) *It fits well* -I feel that the method helped me explain my procedure in a clear and structured way-. 3) *Indifferent* -I don't feel like the method helped or hindered me in my procedure. Each answer had to be justified.

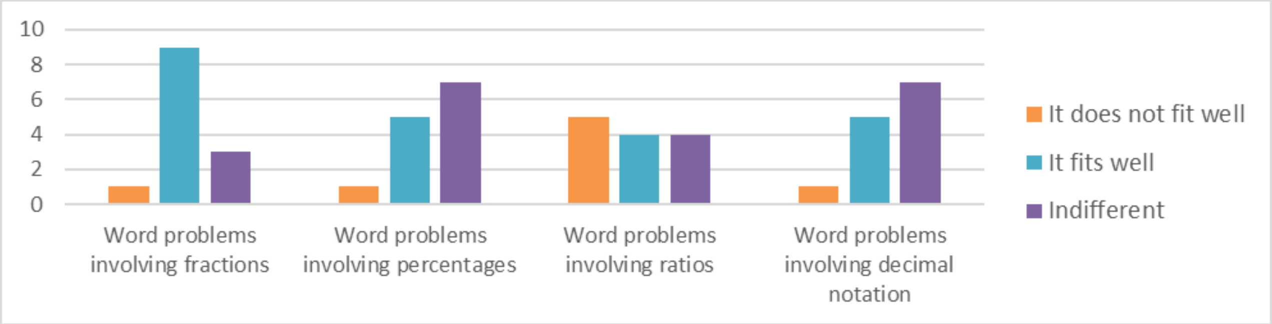


Figure 4: Students' opinion on the use of the method taking into account the content variable

It can be seen in Figure 4 that the mathematical content variable affects the students' perception of how well the method is adapted to capture their thinking process and word problem-solving process. In particular, students think that the analytical readings' method helped them explain their procedures in a clearer way when solving word problems that involve fractions, but this is not the case for those word problems that involve ratios.

Conclusions

The objective of this study was to test the analytical readings' method, regarding its use to assess the processes of word problems solving to document the reflections of a group of students. In this regard, there was evidence that the students considered the analytical readings' method as an assessment tool allows the student's mental activity and thought process to be clearly reflected in their procedure when solving word problems. As each step of the method is related to the phases in the word problem-solving process, it makes easier to follow the student's progress through each of the phases and, if there is an error, determine at what point in the resolution process it occurred, and identify its possible causes. However, according to the students it must be considered that there may be some disadvantages, such as the difficult, and time it takes to use this method when solving a problem.

The use of the analytical readings' method discussed in this paper is as an assessment tool, but during the intervention a topic of discussion among the study participants was the potential of the method as a teaching tool. The advantages of learning to solve certain types of word problems, for example those involving fractions, were discussed. The possibility that this method offers was highlighted so that students can explain their procedures, as well as the development of skills to establish relationships between the quantities that appear in the problem. Although it was also mentioned that using this method for the first time could be a challenge for students.

References

Freudenthal, H. (1983). *Didactical Phenomenology of Mathematical Structures*. Dordrech: D. Reidel.

Polya, G. (1957). How to solve it (2nd ed.). Princeton University Press.

Puig, L. & Cerdán, F. (1988). Problemas aritméticos escolares. Madrid: Síntesis.

Puig, L., & Cerdán, F. (2014). Acerca de carácter aritmético o algebraico de los problemas verbales. En B. Gómez & L. Puig (Eds.), *Resolver problemas: Estudios en memoria de Fernando Cerdán* (pp. 21-34). Valencia: Universitat de València.

Rodríguez, M. A. (2012). Resolución de Problemas. En M. D. Pochulu & M. A. Rodríguez (Eds.), *Educación Matemática: aportes a la formación docente desde distintos enfoques teóricos* (1ra ed., Vol. 1, pp. 153-174). Editorial Universitaria de Villa María, Universidad Nacional de Villa María; Universidad Nacional de General Sarmiento.

Roy, S., & Roth, D. (2015). Solving General Arithmetic Word Problems. En L. Màrquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1743–1752). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1202.

Siegler, R. S., Fazio, L. K., Bailey, D. H., & Zhou, X. (2013). Fractions: The new frontier for theories of numerical development. *Trends in Cognitive Sciences, 17*(1), 13-19. https://doi.org/10.1016/j.tics.2012.11.004

Valenzuela, C. (2018). *Modelo de enseñanza para fracciones basado en la recta numérica y el uso de applets: estudio en comunidades marginadas* [Tesis doctoral]. CINVESTAV.

Verschaffel, L., Schukajlow, S., Star, J., & Van Dooren, W. (2020). Word problems in mathematics education: a survey. *ZDM - Mathematics Education*, 52, 1–16. https://doi.org/10.1007/s11858-020-01130-4

Exploring formative assessment and peer feedback in technology-enhanced mathematics learning environments using bar model virtual manipulatives

Poh Hwee Sim, Claire

Charles University, Faculty of Education, Prague, Czech Republic; clairepohhs@gmail.com

The uptick in the adoption of digital assessment, driven by increased technology integration in classrooms, not only transforms the assessment approach but also holds crucial implications for how teachers assign tasks and shape the way students engage in mathematical reasoning. This paper explores how educators leverage technology to enhance mathematics assessment and feedback. Using screencast (or screen recording) as a primary method of data collection, interactions of students solving word problems utilising the bar model, a web-based virtual manipulative, are recorded. Analysis of data collected may offer insights into students' specific competencies and deficiencies and inform teaching practices to meet their' learning needs. Digital assessment is broadened to include peer feedback and self-evaluation, facilitated by real-time interaction and idea-sharing through screen mirroring, another innovation supported by classroom connectivity.

Keywords: Digital assessment, peer feedback, bar model.

Introduction

Technology has significantly impacted the assessment and feedback processes in teaching mathematics for both summative and formative purposes. The shift from offline to digital assessment introduces fresh possibilities for evaluating mathematics learning and modifying task structures as well as the scope of assessed abilities and skills (Drijvers et al., 2016). The concept of a connected classroom where teachers and students can exchange digital information has persisted for many years (Stacey & Wiliam, 2013). In recent years, technological advancements have rendered this vision more attainable. In the study by Clark-Wilson (2010), using classroom aggregation technology for mathematics was found to have promoted peer assessment as well as self-evaluation. Moreover, it was observed that teachers used feedback from students to inform the planning of future activities. These findings show that classroom connectivity offered fresh possibilities for formative assessment, providing teachers with insights into students' mathematical thinking. Leveraging classroom connectivity, multiple devices may be screen mirrored on the class display for side-by-side comparison and discussion among students. Studies suggest that students adjust their responses when comparing their work with peers, fostering increased opportunities for peer assessment and self-evaluation (Stacey & Wiliam, 2013). Drijvers et al. (2016) outline two crucial steps in formative assessment: collecting data on student achievements and devising strategies to enhance performance. Screen recordings serve as primary data. Teachers review these recordings and formulate appropriate measures to improve performance.

Theoretical Background

Framework for technology-mediated feedback

Mayer's (2002) Cognitive Theory of Multimedia Learning posits that learning is enhanced when information is presented through multiple modalities, such as visual and auditory channels. Figure 1 illustrates a cognitive theory of multimedia learning. Mayer asserts that multimedia messages that engage these cognitive processes are more likely to promote meaningful learning. His findings support a social agency extension of the cognitive theory of multimedia learning, suggesting that social cues within multimedia messages activate a conversational schema in learners, prompting deeper cognitive engagement. This holds significant implications for how the theory will shape peer feedback and assessment practices. This approach resonates with established conversational theories like Grice's (1975) conversational norms, which underscore the dedication to comprehending the other speaker's communication. Two approaches exist for assessing learning: retention tests and transfer tests (Mayer & Wittrock, 1996). Retention tests assess the capacity for memory recall. Transfer tests evaluate how effectively learned knowledge is applied to novel situations. He asserts that transfer tests offer the best assessment of learner understanding. Mayer's hypothesis proposes that better transfer is facilitated through interaction and conversation.

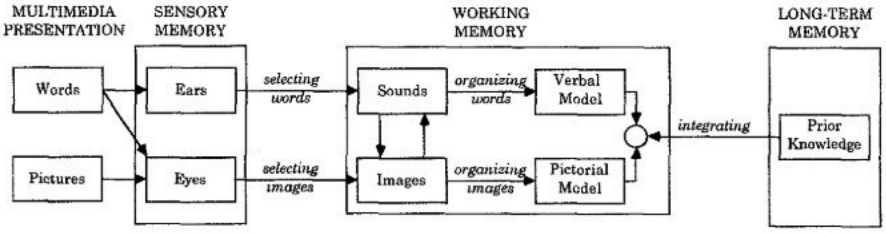


Figure 1: A cognitive theory of multimedia learning (Mayer, 2002, p. 103)

Findings suggest that side-by-side comparison techniques, such as Visual Analysis for Image Comparison (VAICo), offer advantages in terms of speed, clarity, and accuracy in identifying differences in image data (Schmidt et al., 2013). Figure 2 shows various image sets from diverse domains, showcasing the method's adaptability across datasets. A crucial aspect of effective data analysis involves selecting suitable similarity metrics. We suggest teachers adopt this selection method when choosing screenshots of students' solution for side-by-side screen comparison to teach bar model strategies. Rittle-Johnson et al. (2017) underscore the significance of comparison in conceptual learning, with their classroom-based research supporting its efficacy in algebra instruction. Mayer (2002) suggests expanding the cognitive theory of multimedia learning to include social factors affecting learners' engagement in deep cognitive processing, such as combining visual (e.g., selective screenshots) and verbal (e.g., feedback interaction) models. Building on the body of research, we argue that when teachers present students' solutions using side-by-side screen for comparison, they foster real-time sharing and collaboration among students, contributing to multimedia learning through peer feedback.

Framework for the model method

A distinctive pedagogy of Singapore mathematics, the model method is inspired by Greeno's part-whole and comparison schemas (Nesher, Greeno & Riley, 1982; Kintsch & Greeno, 1985). Students use rectangular bars to visualize mathematical relationships, facilitating comprehension of abstract quantities. For discussion in this study, consider the following illustrations of part-whole and comparison models.

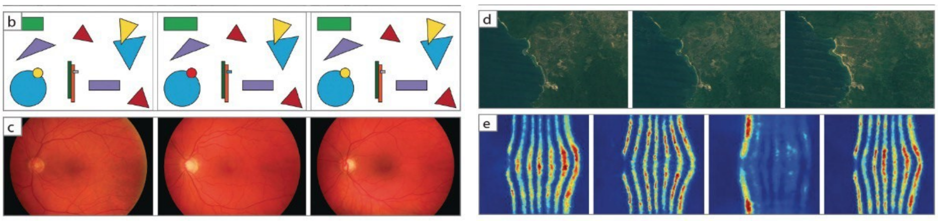
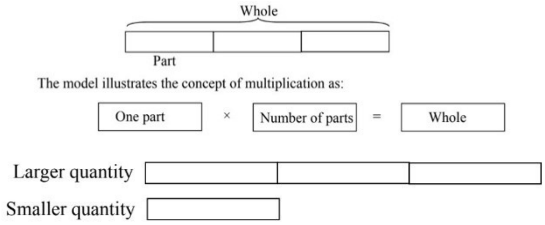


Figure 2: Image datasets (Schmidt et al., 2013, p. 6) (b) shapes disappear, re-appear or change their colour (c) retina images from different patients (d) satellite images of a coastline affected by tsunami in Indonesia (e) images with colour coded gene expression information

Table 1: Part-whole model for multiplication and division & multiplicative comparison models (Kho et al., 2014, p. 227)

 Whole Part The model illustrates the concept of multiplication as: One part × Number of parts = Whole Larger quantity Smaller quantity	Part-whole model: The total is determined by the multiplication of one part and the number of parts. Conversely, if we know the total and one factor, we can find the other factor through division. Comparison model: The larger quantity is three times the smaller quantity, and conversely, the smaller quantity is one-third of the larger quantity. For example, if the larger quantity represents three units, the smaller quantity represents one unit. Together, they total four units, with a difference of two units between them.

Research Questions

The study seeks to examine digital assessment and feedback through an extension of mathematical experience using integrated technology. The research questions guiding the study are as follows:

RQ1. How can teachers leverage classroom connectivity to effectively analyse students' conceptual deficiencies in word problem solving utilising bar model virtual manipulatives?

RQ2. How can side-by-side screens be utilised for peer feedback in a technology-enhanced classroom?

Methods

This study analysed students' digital experiences within technology-enhanced mathematics learning environments, focusing on formative assessment and feedback. For this research, a programme was piloted with an elementary school in the Czech Republic involving a cohort of nine Grade 8 (age 14) participants, during which they engaged in word-problem solving, utilising bar model virtual manipulatives apps provided within their tablet devices. The study leveraged classroom connectivity through digitally accessing and analysing students' mathematical thinking via screencast, i.e., recordings of their on-screen interactions while using bar model manipulatives. The students were given a series of tasks during which on-screen activities were recorded. The pilot programme, led by the researcher who also served as the teacher, consisted of two 3-hour sessions, one-week apart.

Data collection and analysis

Screencast was utilised as the primary means of collecting data, capturing visual information of students' digital interactions. Following data collection, the teacher-researcher commenced the analysis process by describing and interpreting visual cues within the datasets in the form of analytic memos. The following sections outline the data collection process, which generated samples for analysis. This is followed by sorting and selecting samples for class discussion utilising side-by-side screen. Figure 3 is a flowchart illustrating this process.

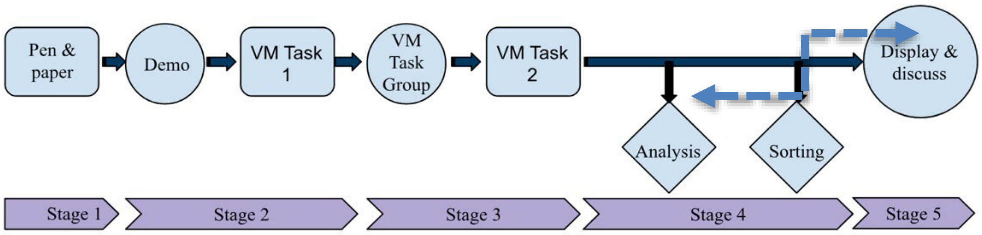


FIGURE 3: INSTRUMENTATION, DATA COLLECTION AND ANALYSIS

Stage 1: Pen and paper task – Baseline establishment

Stage 2: Demo and Task 1 – Standardised demo and Task with virtual manipulatives

Stage 3: Group activities (VM Task Group) and Task 2 (VM Task 2) - Participants collaboratively solved word problems in groups of four or five. The aim was to assess the impact of group activities on students and determine if peer feedback is evident thereafter. Task 1, Group Activities and Task 2 were recorded for analysis, concluding Day 1 of the study.

Stage 4: Analysis and sorting - This involved systematically reviewing screen recordings of Task 1, Task 2 and Group Activities and writing analytic memos to capture emerging themes and pattern:

Initial Observations – The recordings were viewed multiple times for initial understanding.

Identifying Patterns – Patterns, themes and significant moments in the interactions were identified (see Coding). The identification process also included selecting among models with correct solution and incorrect solutions and sorted according to similarity metrics (Schmidt et al., 2013).

Coding – Codes were applied to segments of the screen output to represent moments of interactions. Interactions contributing to conceptual understanding: 1) use of the appropriate concept (Table 1), 2) accuracy in partitioning 3) alignments of parts 4) correct labelling, 5) application of operations

Memo Writing – Specific moments in the videos were annotated, documenting observations, interpretations and insights corresponding to coded segments.

Interpretation I – Coded segments and memos were analysed to identify broader themes related to research objectives. For RQ 1, data on interactions highlighting both strengths and weaknesses in the topic was collected and interpreted. Next steps were formulated to address weaknesses.

Stage 5: Display and discuss

Interpretation II – The selected screenshots served as instructional material and were displayed in class for side-by-side comparison and discussion. Observations were documented and redirected to Stage 4 for analysis (Figure 2). For RQ 2, reflective prompts (cognitive transfer and image comparison) guided reflections on students' engagement and comments. Cognitive transfer focused on identifying their ability to apply learned concepts in new situations. Image comparison related to their perception of accuracies/inconsistencies in screenshots and observed cross-referencing and editing of their own model construction.

Results

Interpretation I

This section addresses RQ1, examining interactions to identify strengths and weaknesses in the topic. We analyse Jakub's on-screen data to understand his grasp of the word problem in detail.

Question: 3/5 of the students in Grade 8 and 2/3 of the students in Grade 7 are girls. Both classes have the same number of girls. Grade 8 has 4 more boys than Grade 7. How many students are there in Grade 8?

For clarity, sequential screenshots (Transformations) of significant interactions are provided (Table 2b) with analytic notes for each moment. Table 2a outlines interpretations and reflections for these interactions. The documentation includes: 1) Appropriate Concept: Yes 2) Partitioning Accuracy: Yes 3) Alignment of Parts: No 4) Correct Labelling: No 5) Application of Operations: No.

The applied codes effectively pinpoint locations in the video, facilitating analysis. A thorough screencast examination reveals precise areas where Jakub struggled, with analytical notes explaining underlying reasons. Despite using the appropriate model, Jakub's solution was incorrect. The on-screen data not only clarifies specific points of struggle but also underscores his potential to engage with a concept he correctly selected but has not yet mastered in its application. Tailored remediation strategies can address Jakub's conceptual gaps, guiding future instructional strategies.

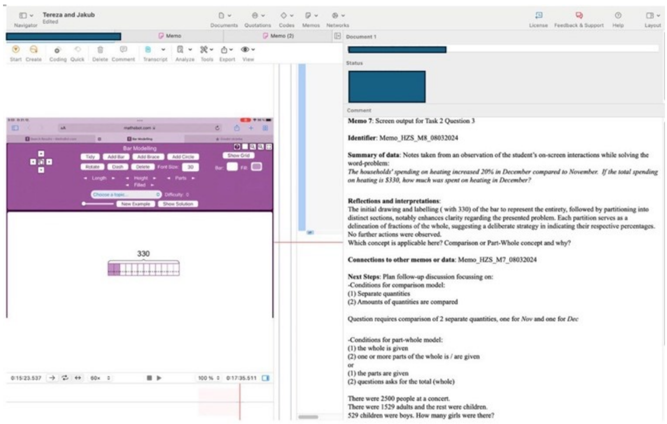
Interpretation II

We use screenshots from Tereza's and Jakub's screen output (Figure 5 and Figure 6) to address the research objective that side-by-side screen effectively support peer feedback in a technology-enhanced classroom. The screenshots were selected using similarity metrics, aligning with the notion that selecting suitable similarity metrics is crucial for effective data analysis (Schmidt et al., 2013). This adds validity to their utilization in the study. Exchanges between teachers and students primarily reflected prompts and interactions related to the word problem. No individual attribution was recorded regarding which participant commented or posed each question during the study.

Table 2a: Reflections and interpretations of significant interactions

(a) The visual appears to be an effort to understand whether the number of girls aligns with the specified conditions outlined in the given word problem.
(b) The transformation shows that the appropriate bar model concept has been applied, i.e. the comparison model. This grasp of selecting the suitable application of the concept may potentially be attributed to the peer collaboration or peer learning experienced during the preceding group activity.
(c) The consistent use of the 'x' notation across all units led to confusion, resulted in him reaching an impasse.
(d) This visual representation highlights a discrepancy in logic, where the number of units drawn is unequal despite being depicted in the same size.

In addressing how side-by-side screens are utilised for peer feedback, we integrate Mayer's (2002) method of employing transfer tests to assess learner understanding. Questions and comments are famed and analysed from the perspective of cognitive transfer. Furthermore, we utilise similarity metrics to improve the effectiveness of data analysis (Schmidt et al., 2013) and to foreground inconsistencies between bar model constructions. When students identify these inconsistencies, peer feedback is engaged. Details of the analysis are provided in Table 3a and Table 3b.



Memo 7: Screen output for Task 2 Question 3

Identifier: Memo_HZS_M8_08032024

Summary of data: Notes taken from an observation of the student's on-screen interactions while solving the word-problem:
The households' spending on heating increased 20% in December compared to November. If the total spending on heating is $330, how much was spent on heating in December?

Reflections and interpretations:
The initial drawing and labelling (with 330) of the bar to represent the entirety, followed by partitioning into distinct sections, notably enhances clarity regarding the presented problem. Each partition serves as a delineation of fractions of the whole, suggesting a deliberate strategy in indicating their respective percentages. No further actions were observed.
Which concept is applicable here? Comparison or Part-Whole concept and why?

Connections to other memos or data: Memo_HZS_M7_08032024

Next Steps: Plan follow-up discussion focussing on:
-Conditions for comparison model:
(1) Separate quantities
(2) Amounts of quantities are compared

Question requires comparison of 2 separate quantities, one for *Nov* and one for *Dec*

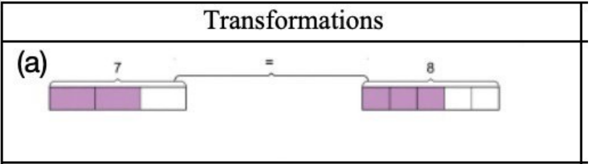
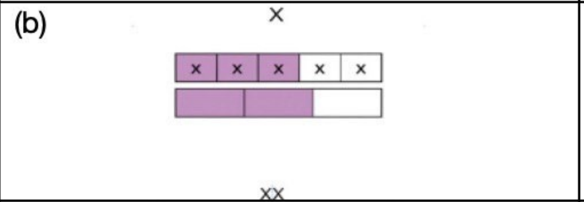
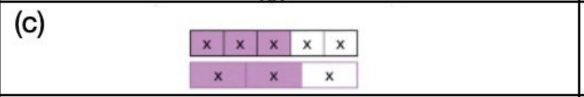
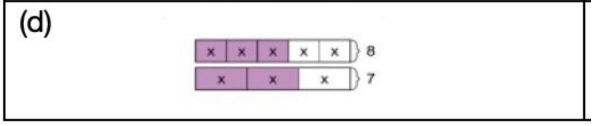
-Conditions for part-whole model:
(1) the whole is given
(2) one or more parts of the whole is / are given
or
(1) the parts are given
(2) questions asks for the total (whole)

There were 2500 people at a concert.
There were 1529 adults and the rest were children.
529 children were boys. How many girls were there?

Figure 4: Jakub's data and analytic memo

This illustrates on-screen video data uploaded for analysis using qualitative research tool. The qualitative tool utilised enables thorough capturing of significant interactions and facilitates the creation of corresponding analytic memos

Table 2b: Sequential screenshots of Jakub's data

Transformations





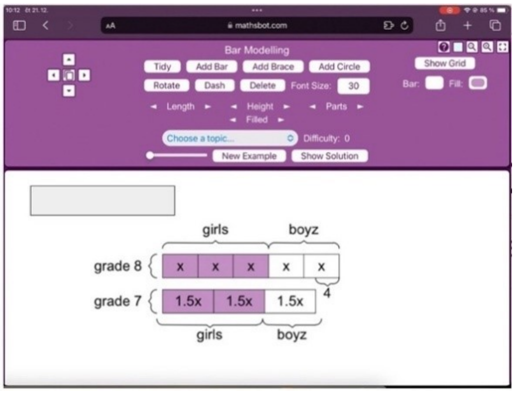


Figure 5: Tereza's screenshot

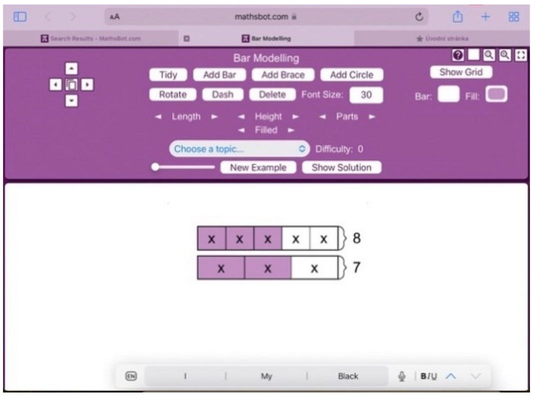


Figure 6: Jakub's screenshot

Note. From MathsBot.com (Hall, 2013)

Table 3a: Analysis of students' contributions

Student's contributions:	Cognitive Transfer	Image Comparison
That (pointing to Jakub's model) is not correct.	By identifying errors or inconsistencies in peer's work, the student demonstrates understanding of the subject matter and capacity to critically analyse problem-solving approaches.	Further discussion about why he/ she thought it was incorrect triggers a number of debates among the students. Some were not agreeable there were any errors, others asserted that there were errors which prompted the comments that follow.
5x is not equal to 3x	This observation indicated the student's ability to apply his/ her understanding of mathematical concepts, specifically the relationship between quantities represented by variables (5x and 3x)	This comment shifted some students' attention to Tereza's model, with one pointing out the shaded regions indicated equal values. Someone brought up this might mean Tereza's model was wrong, since 5x was not equal to 3x. Several others disagreed.
One bar should be longer and one bar should be shorter	This statement reflected the student's observation of a discrepancy in the lengths of bars depicted in a diagram, showcasing his/ her ability to apply mathematical concepts to evaluate visual information presented in the diagram.	There were some initial confusion which model or bar was being referred to. This led to a deliberation that concluded that one class has 4 more boys than the other.. A number of students started turning to their devices to rework their constructions, others were seen cross-referencing each other's model

Table 3b: Analysis of teacher's contributions

Teacher's contributions:		
There are the same number of girls in both classes (restating the important information).	Highlighting key words to restate the problem prompting students to re-address the problem in a structured way. It brought focus to the problem-solving transfer, promoting meaningful learning.	Several students started to point out the same number of girls were represented by two perfectly aligned shaded parts on Tereza's models.
How do we show that 3/5 of the students in Grade 8 and 2/3 of the students in Grade 7 have the same value? Show this on your bar model.	This is a critical juncture in the learning process. The students' attention was drawn to a seemingly different but equivalent values. This is an opportunity for meaningful learning as they grapple with complex ideas and develop strategies	Some students noticed that Tereza's model fitted the description. A number of them cross-referenced their own models with Teresa's model and were making changes.
Everyone, draw (the bars) on your screen. How do you make the rectangles equal?	This was an opportunity to be seized upon as the bar model approach lends itself to effectively convey abstract concepts through its visual representation.	There were significantly more discussions, some revising their constructions, shading, re-labelling to emphasise this aspect of the information discussed

They were then asked to write down the algebraic equation. Eventually, they arrived at the equation: $5x - 4 = 4.5\ x$

They then solved for x by transposition and arrived at the answer: $x = 8$; Answer 40

Discussion and initial findings

Drijvers et al. (2016) resonate with our study's findings, indicating a shift in mathematics pedagogical practices from teacher-led demonstrations to student-led modelling and discussions facilitated by technology integration. The analysis outcome satisfies the research goals of using technology to help teachers effectively analyse students' conceptual gaps. Screencasts enabled the teacher to monitor students' math activity unobtrusively, facilitating authentic feedback to offer tailored support more effectively. The gradual capture of Jakub's screen output showcased his potential capacity in applying a concept, an aspect that would not be evident if viewed solely as a finished product, e.g., on pen and paper or static display. Further, the analysis suggests that side-by-side screen shows promise in supporting peer feedback. The images featuring similarity metrics sparked discussions among the students, fostering cognitive transfer as they exchange perspectives while examining the images they are comparing.

Conclusion

While the arguments presented strongly advocate for utilising digital means for assessing mathematics, in practice, implementing the features discussed remains challenging. In our research, we recognised a constraint regarding the efficiency of using screencasts for assessment, particularly in larger classroom settings. While our study involves a modest sample size of 9 students, the scalability of screencasts becomes challenging when applied to classrooms with more students. Managing, reviewing, and evaluating a large volume of screencasts poses logistical and practical hurdles for educators. This limitation highlights the need to explore alternative assessment strategies or technological solutions to ensure effective assessment practices in larger classroom environments. Another significant constraint is the lack of robust tools to cater for authentic mathematical practices such as sketching and scribbling within the app. Certainly, resorting to paper and pen can circumvent these constraints. However, this approach proves impractical when the objective is assessment through technology, as only a portion of the student's work would be visible within the assessment system. As a result, students may struggle to demonstrate their full problem-solving abilities, leading to a misalignment between their mathematical competence and the assessment's practice domain (Drijvers et al., 2016).

References

Clark-Wilson, A. (2010). Emergent pedagogies and the changing role of the teacher in the TI-Nspire Navigator-networked mathematics classroom. *ZDM–The International Journal of Mathematics Education, 42*(7), 747–761.

Drijvers, P., Ball, L., Barzel, B., Heid, M. K., Cao, Y., & Maschietto, M. (2016). *Uses of Technology in Lower Secondary Mathematics Education: A Concise Topical Survey*. ICME-13 Topical Surveys, 1–34. http://doi.org/10.1007/978-3-319-33666-4

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts* (pp. 51–58). Brill. https://doi.org/10.1163/9789004368811_003

Hall, J. (2013). *MathsBot.com*. https://mathsbot.com/

Kho, T. H., Yeo, S. M., & Fan, L. (2014). Model method in Singapore primary mathematics textbooks. In K. Jones, C. Bokhove, G. Howson, & L. Fan (Eds.), *Proceedings of the International Conference on Mathematics Textbook Research and Development* (pp. 275–282). University of Southampton.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving arithmetic word problems. *Psychological Review, 92*(1), 109–129. https://doi.org/10.1037/0033-295X.92.1.109

Mayer, R. E. (2002). Multimedia learning. *Psychology of Learning and Motivation, 41*, 85–139. https://doi.org/10.1016/S0079-7421(02)80005-6

Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. In D. Berliner and R. Calfee (Eds.), *Handbook of educational psychology* (pp. 45–61). Macmillan.

Nesher, P., Greeno, J. G., & Riley, M. S. (1982). The development of semantic categories for addition and subtraction. *Educational Studies in Mathematics, 13*(4), 373–394. https://doi.org/10.1007/BF00366618

Rittle-Johnson, B., Star, J. R., & Durkin, K. (2017). The power of comparison in mathematics instruction: Experimental evidence from classrooms. In D. C. Geary, D. B. Berch, and K. M. Koepke (Eds.), *Acquisition of complex arithmetic skills and higher-order mathematics concepts* (pp. 273–295). Elsevier. https://doi.org/10.1016/B978-0-12-805086-6.00012-6

Schmidt, J., Gröller, M. E., & Bruckner, S. (2013). VAICo: Visual Analysis for Image Comparison. *IEEE Transactions on Visualization and Computer Graphics, 19*(12), 2090–2099. https://doi.org/10.1109/TVCG.2013.213

Stacey, K., & Wiliam, D. (2012). Technology and assessment in mathematics. In M. A. Clements, A. J. Bishop, C. Keitel, J. Kilpatrick & F. K. S. Leung (Eds.), *Third international handbook of mathematics education* (pp. 721–751). Springer. https://doi.org/10.1007/978-1-4614-4684-2_23

Conception of two informative tutoring feedback strategies for mathematical tasks with STACK

Farhad Razeghpour

Ruhr-University Bochum, Germany; farhad.razeghpour@ruhr-uni-bochum.de

The increasing prevalence of computer-based teaching and learning opportunities in recent years has led to the emergence of novel approaches for providing informative tutoring feedback (ITF). For mathematical tasks, the tool STACK represents a promising avenue for exploration. STACK enables the automated evaluation of open mathematical tasks and the provision of immediate feedback for learners. This paper will therefore investigate the potential of STACK in designing ITF-strategies for mathematical tasks. In light of this question, two distinct strategies are presented, which differ in the case of incorrect answers with unclear causes. While in one strategy learners are then provided with solution-specific hints for processing the task, in the other strategy they enter a task loop in which they must work out these hints on their own. Finally, the strategies are compared in terms of their characteristics and possible effects on students' learning processes.

Keywords: Computer assisted learning, education, informative tutoring feedback, mathematics.

Introduction

Feedback plays a pivotal role in mathematics education, exerting a profound influence on enhanced academic performance, with a considerable impact on student learning (Hattie, 2010). When designing feedback, it is essential to determine whether it is incorporated into a formative or summative assessment. Summative assessments are provided at the end of the learning process and serve to evaluate the learning outcomes. In contrast, formative assessments are provided during the learning process with the intention of having a positive impact on it (Pals et al., 2023). This paper focuses on formative assessments as the objective is to design feedback that supports learning processes.

In recent years, computer-based teaching and learning tools have offered new ways of providing informative tutoring feedback (e.g. Erickson et al., 2020). This refers to feedback for digital learning environments that provide learners with solution- and error-specific hints rather than concrete solutions, enabling them to correct their incorrect answers independently. In the field of mathematics teaching, the STACK tool represents a promising approach for this. STACK enables the automatic assessment of open mathematical tasks and provides learners with immediate feedback (Knaut et al., 2022).

This theoretical paper aims to investigate the use of STACK for the realization of effective ITF-strategies for mathematical tasks. The theoretical background highlights the central role of STACK in automatizing the assessment process for mathematical tasks. With its integrated computer algebra system (CAS), STACK enables the evaluation of students' answers to various mathematical tasks. Furthermore, this section describes the relevant characteristics of ITF-strategies. Subsequently, the research question of how STACK could be used to design ITF-strategies for mathematical tasks is addressed. To answer this research question, two different implementations of ITF-strategies are presented and explained. The effectiveness of both strategies is discussed on a theoretical basis, highlighting their potential impact on students' learning processes. The paper concludes with a summary and an outlook on further related issues.

Theoretical background

STACK

Computer-based tools can be utilized to automatically evaluate tasks with even free input fields and to provide learners with feedback concerning their achievement. For mathematical tasks, this can be accomplished with STACK (System for Teaching and Assessment using a Computer algebra Kernel). Due to the implemented CAS, the digital evaluation of tasks with any number of correct solutions is possible (Alarfaj & Sangwin, 2022). Furthermore, emerging numerical values can be randomized. As a result, learners are offered numerous opportunities to practise by creating a single task. Meanwhile, STACK is available as a free plugin for Moodle, allowing the creation and use of STACK tasks in this learning platform.

For the digital evaluation and provision of feedback, a so-called response tree must be constructed in advance for each task (see Figure 1). Within the response tree, the learners' answers are checked for various mathematical characteristics in several nodes (Knaut et al., 2022). Through the underlying CAS, it is possible, for example, to check whether an entered function has certain zeros or extreme points. After a positive or negative result of the check, the path can be stopped or connected to another node for further analysis. In Figure 1, the paths are visualized by green and red lines. On each path, feedback can be stored and points for performance can be added or subtracted. Overall, the construction of a response tree enables the evaluation of various answers and the provision of feedback for learners (Sangwin, 2023).

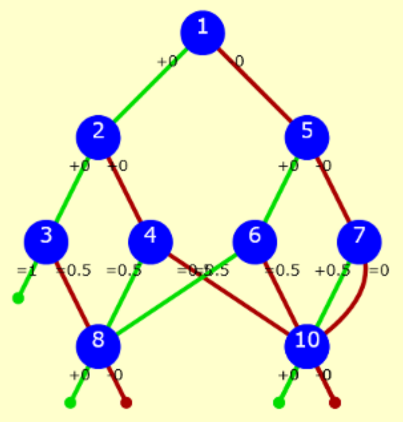


Figure 3: Example of a response tree for a task

Informative tutoring feedback strategies

Narciss (2008) has classified various forms of feedback in terms of their content. A distinction is made between simple and elaborated forms. While simple forms of feedback only include information about the correctness of a task, elaborated forms include more extensive information. Examples of simple feedback are *knowledge of performance* (KP), which indicates the number or proportion of tasks solved correctly, and *knowledge of result* (KR), which contains a specification of which tasks were solved correctly. An example of an elaborated form is feedback that discusses specific errors and offers explanations for their occurrence, which is assigned to the category knowledge about errors (KM). If solution-specific hints are included in the feedback, it is referred to as knowledge on how to proceed (KH). Solution-specific hints contain information about strategies and necessary intermediate steps to solve the task without presenting a complete solution.

In the case of digital learning tools such as the STACK system previously presented, a feedback strategy can be employed to define which feedback forms are combined and presented to learners. Accordingly, a feedback strategy is a defined plan that determines the structure and presentation of feedback. One approach is to develop ITF-strategies, which combine simple KR-feedback with elaborated forms such as KH- and KM-feedback (Narciss, 2012). Instead of providing learners with correct answers, the focus is on enabling them to use the feedback to correct their responses. As a result, learners should be allowed to retry tasks immediately after receiving feedback. The opportunity for immediate implementation increases the relevance of the feedback for the learners and leads to a more intensive engagement with it (Tärning et al., 2020). This is intended to facilitate the students' learning processes.

Research question

In light of the potential for ITF-strategies to support learning processes, it is necessary to consider how these strategies could be implemented in practice with current technologies. While there have been a few studies on the implementation of ITF-strategies in digital learning environments for specific mathematical areas, such as written subtraction (Narciss & Huth, 2004), further research is needed on how ITF-strategies could be implemented in general for mathematical tasks. This research is essential for the wider use of these strategies in mathematics classes and for the investigation of them in experimental studies. The STACK system could provide a possible approach to this need. This paper will therefore investigate the research question of how STACK could be used to design ITF-strategies for mathematical tasks. To answer the research question, two possible approaches using STACK are developed, which both respect the presented characteristics of ITF-strategies. The two strategies are described in detail in the subsequent chapter.

Conception of ITF-strategies with STACK

Firstly, we consider the two cases in which either a correct solution or an erroneous solution is entered where the cause of the error is recognizable. The design of the two strategies is identical for these two cases. In both strategies, students receive KR-feedback in the event of a correct input. The correct solution can be verified in STACK in two different ways due to the underlying CAS. The answer can be compared with a saved sample solution for algebraic equivalence or tested for relevant properties. These options ensure that every correct answer is recognized, even if it is entered in a different form (e.g. fraction instead of a decimal number) or if there are any number of correct answers. For certain incorrect answers, both strategies offer KM feedback. This necessitates the implementation of appropriate checks in advance within the response trees. Due to the CAS, it is sufficient to test the submitted answer for relevant mathematical characteristics and hence no explicit incorrect solutions need to be stored within the response trees. For example, it is possible to ascertain whether an inserted function has the requisite degree or whether the correct variable has been used. This approach enables the identification of numerous incorrect responses, all attributable to a common set of error sources, within individual nodes. In this way, differentiated KM-feedback can be formulated for several typical errors.

The described common components of both strategies are illustrated in Figure 2 in the form of a diagram for an exemplary integral task. As previously stated, KR-feedback is displayed if the input is correct. In the event of an incorrect answer, several different error-specific hints have been stored where the cause of the error can be deduced reliably enough to provide KM-feedback. For the sake of clarity, not all five KM-feedback cases are listed in full in Figure 2, but only two of them. In the first case shown (see top left in Figure 2), feedback is given if a solution has been entered that arises when the product of two integrals has been calculated instead of the integral of the product. This scenario utilizes an example to explain why this method is typically not allowed. The second case, depicted in the top right of Figure 2, concerns a scenario where the correct approach of partial integration has been selected, but the terms in the formula have been added instead of subtracted. The formula is shown, with the minus sign highlighted in red.

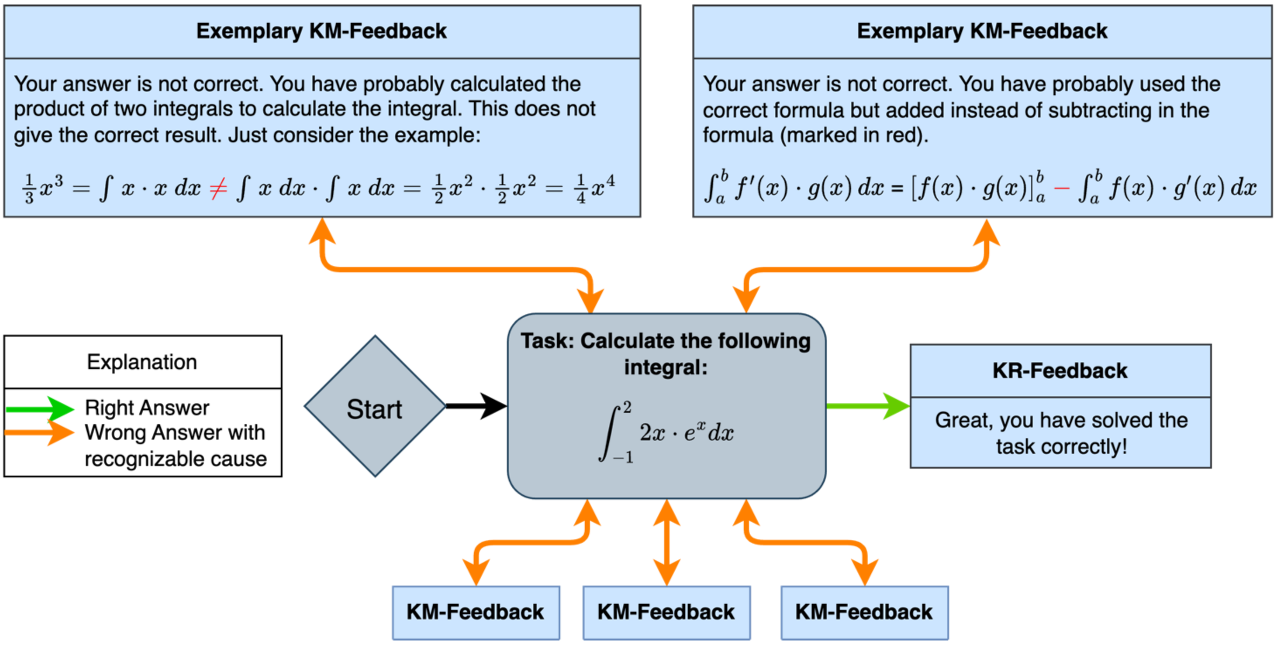

Exemplary KM-Feedback
Your answer is not correct. You have probably calculated the product of two integrals to calculate the integral. This does not give the correct result. Just consider the example:
$\frac{1}{3}x^3 = \int x \cdot x\, dx \neq \int x\, dx \cdot \int x\, dx = \frac{1}{2}x^2 \cdot \frac{1}{2}x^2 = \frac{1}{4}x^4$
Exemplary KM-Feedback
Your answer is not correct. You have probably used the correct formula but added instead of subtracting in the formula (marked in red).
$\int_a^b f'(x) \cdot g(x)\, dx = [f(x) \cdot g(x)]_a^b - \int_a^b f(x) \cdot g'(x)\, dx$
Explanation
Right Answer
Wrong Answer with recognizable cause
Start
Task: Calculate the following integral:
$\int_{-1}^{2} 2x \cdot e^x dx$
KR-Feedback
Great, you have solved the task correctly!
KM-Feedback
KM-Feedback
KM-Feedback

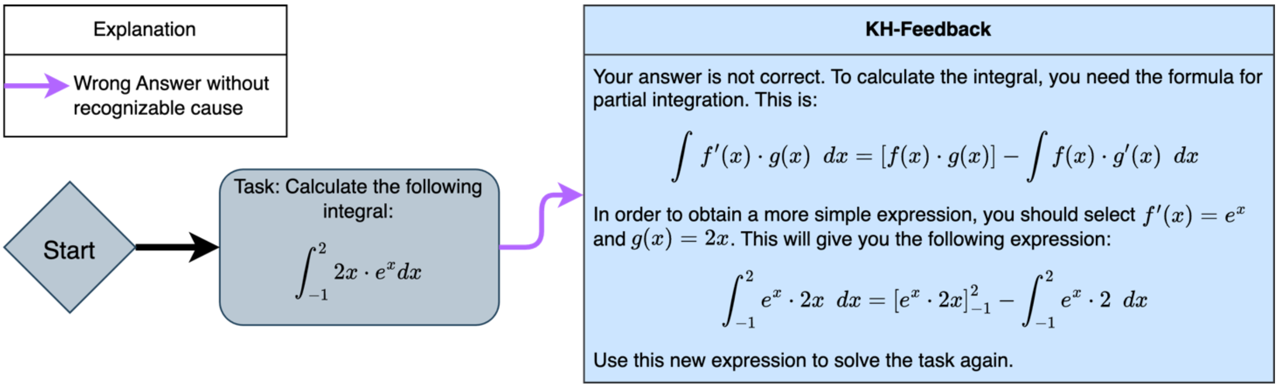
Figure 2: Common feedback components of both strategies in an integral task

Now the challenging question arises of how to deal with certain incorrect entries for which no specific error causes have been created in the response tree. This is a likely scenario, as task creators cannot consider every possible cause of error in advance. In addition, particularly in the case of more complex tasks, a combination of several error causes can result in incorrect answers that cannot be recognized based on the final solution. Nevertheless, it is of the utmost importance to ensure that students receive constructive feedback even in the event of such erroneous entries, in order to enable them to identify their mistakes and rectify their solutions. The two feedback strategies deal with this situation differently.

As part of the first feedback strategy, students receive KH-feedback in the event of an incorrect answer that was not saved in the response tree. The feedback contains solution-specific instructions for processing the task. However, the correct solution is not anticipated, as the students are asked to use the information presented to perform the last step independently. In this way, learners still have the opportunity to correct their answers. We describe this strategy as *summarizing* feedback.

This case is illustrated in Figure 3 for the integral task. In the event that an incorrect response cannot be attributed to an underlying error, we present the formula for partial integration in the KH-feedback. Furthermore, the choice of functions for the two factors and the resulting expression are presented. It is now the learner's task to calculate this new integral expression.

However, the second feedback strategy handles this case differently. In the event of an incorrect answer, for which the cause of the error is not stored in the response tree, the second feedback strategy does not provide any error-specific feedback. Instead, students can enter a task loop in which they


Explanation
Wrong Answer without recognizable cause
Start
Task: Calculate the following integral:
$\int_{-1}^{2} 2x \cdot e^x dx$
KH-Feedback

Your answer is not correct. To calculate the integral, you need the formula for partial integration. This is:

$$\int f'(x) \cdot g(x) \ dx = [f(x) \cdot g(x)] - \int f(x) \cdot g'(x) \ dx$$

In order to obtain a more simple expression, you should select $f'(x) = e^x$ and $g(x) = 2x$. This will give you the following expression:

$$\int_{-1}^{2} e^x \cdot 2x \ dx = [e^x \cdot 2x]_{-1}^{2} - \int_{-1}^{2} e^x \cdot 2 \ dx$$

Use this new expression to solve the task again.

Figure 3: KH-component of the summarizing feedback in an integral task

work their way through the entire task in a series of sub-steps. In these task loops, the solution-specific hints from the summarizing feedback are now reformulated as questions which now should be answered by the students. They also receive error-specific feedback while working through the sub-steps. This feedback is usually more extensive than the feedback received when working through the initial complex task. The reason for this lies in the fact that more precise insights into the causes of errors can be drawn from the answers to the sub-steps. We refer to this strategy as *guiding* feedback.

In Figure 4, the case in which an incorrect response was entered that cannot be traced back to the cause of its error is now visualized for the guiding feedback. Concerning this situation, a button is displayed in the feedback that leads to the task loop (see Figure 4). The first sub-step relates to the necessary method for calculating the integral. In a drop-down menu, the correct solution (partial integration) among some distractors can be selected. If the answer is incorrect, KM-feedback is given for each answer option, explaining why the selected option is not appropriate or not permitted. If the answer to the first sub-step is correct, KR-feedback appears, which leads to the next sub-step. While the second sub-step still has a closed answer format, the third, fourth and fifth sub-steps employ an open input field format. For these three open input fields, a total of ten error-specific hints are provided if required. These error-specific hints are not explicitly shown in Figure 4 for reasons of clarity. In the fifth sub-step, learners are asked to specify a suitable antiderivative. Once this last sub-step has also been successfully completed, students return to the initial task. Subsequently, learners may re-engage with the task using the sub-solutions they have worked out. Their sub-solutions are displayed so that they can easily access them. The task loop described can be repeated if necessary. The creation of these task loops is not a regularly offered feature of STACK. This option was developed at the Ruhr University Bochum by incorporating an additional JavaScript program (Altieri et al., 2020).

It can be argued that both strategies have the potential to support students' learning processes. Both strategies contain at least one important elaborated component in addition to the evaluative one (Kulhavy & Stock, 1989). Moreover, both strategies encourage learners to become active and correct their answers independently. In both cases, this could lead to a more intensive engagement with the feedback and thus have a positive effect on the learning process (Tärning et al., 2020).

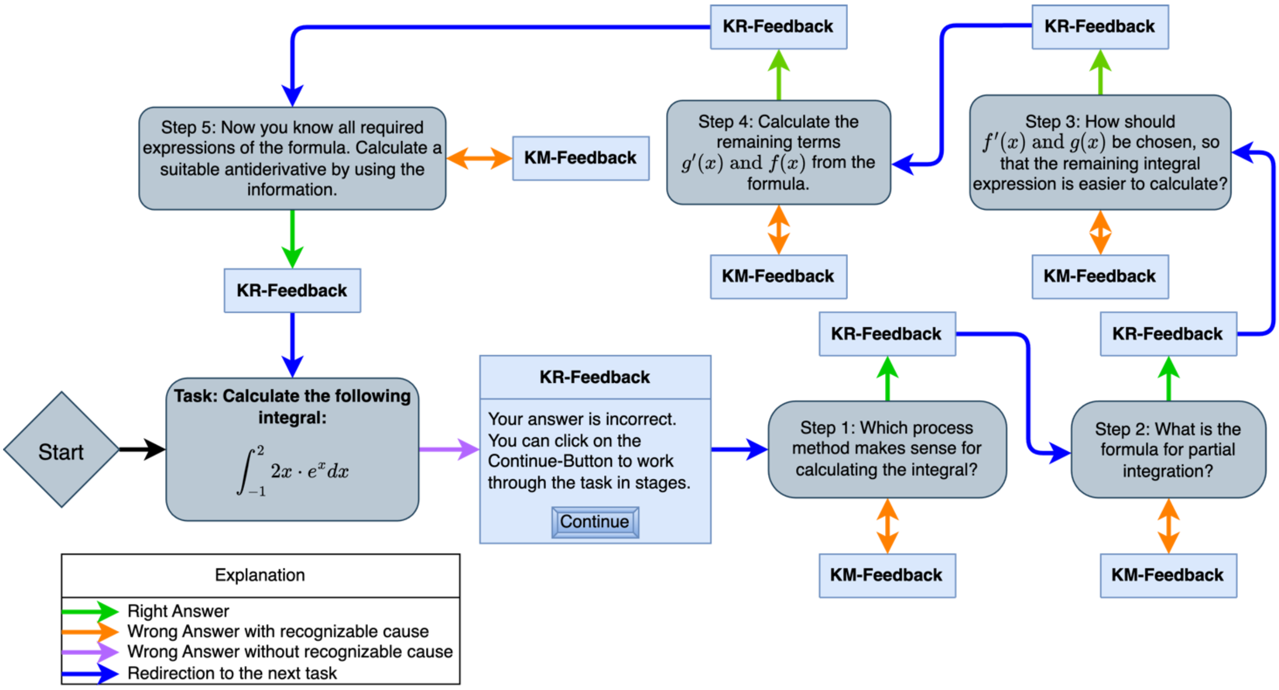

KR-Feedback
KR-Feedback
Step 5: Now you know all required expressions of the formula. Calculate a suitable antiderivative by using the information.
KM-Feedback
Step 4: Calculate the remaining terms $g'(x)$ and $f(x)$ from the formula.
Step 3: How should $f'(x)$ and $g(x)$ be chosen, so that the remaining integral expression is easier to calculate?
KR-Feedback
KM-Feedback
KM-Feedback
KR-Feedback
KR-Feedback
Start
Task: Calculate the following integral:
$\int_{-1}^{2} 2x \cdot e^x dx$
KR-Feedback
Your answer is incorrect. You can click on the Continue-Button to work through the task in stages.
Continue
Step 1: Which process method makes sense for calculating the integral?
Step 2: What is the formula for partial integration?
KM-Feedback
KM-Feedback
Explanation
Right Answer
Wrong Answer with recognizable cause
Wrong Answer without recognizable cause
Redirection to the next task

Figure 4: Task loop of the guiding feedback in an integral task

One favor of summarizing feedback lies in the fact that solution-specific cues target correct solution strategies. Therefore it could have a more motivating effect (Fong et al., 2019). In addition, similar to the use of correct solutions (Renkl, 2014), the presentation of solution-specific hints can be assumed valuable. Solution-specific hints draw students' attention to the correct solution procedure, making it easier for them to focus on acquiring new knowledge (Große & Renkl, 2007). Furthermore, its design could contribute to a clearer overview when working on the task and therefore lead to a better understanding.

Nevertheless, it can be argued that the guiding feedback has a more profound effect on the learning process, as the students are required to comprehend and apply the content of the feedback within the task loops to a greater extent. The individual sub-steps that are displayed in the summarizing feedback must be worked out independently in the tasks with guiding feedback. This could have a direct influence on the self-efficacy of the students, as it provides more opportunities for mastery experiences (Bandura, 1997; Ramdass & Zimmerman, 2008). Another advantage that results directly from the task loops is that errors can be traced back more precisely to their causes when working through the sub-steps. In this manner, the feedback can be employed to ascertain the learning status and assist students in closing learning gaps (Hattie & Timperley, 2007).

Conclusion and outlook

The presented tool STACK is appropriate for the creation of tasks with formative feedback due to its numerous technical possibilities. The response tree in STACK can be used to provide learners with elaborated feedback on their answers. For this purpose, task creators must consider in advance, based on literature or empirical observations, which mistakes learners might make. They can then create response trees with appropriate checks and corresponding KM-feedback. However, the causes of errors can only be anticipated in advance to a certain extent. To also handle incorrect answers that cannot be traced back to specific errors by defined response trees, two feedback strategies were presented in this paper. In the summarizing feedback, solution-specific hints are presented. Despite this, in the guiding feedback, only error-specific hints are given, while intermediate solutions must be worked out in sub-steps.

The question of whether one of these feedback strategies offers greater support for students' learning processes requires empirical investigation. On the basis of theoretical considerations, reasons could be identified for both feedback strategies explaining their potential support for learning processes. Comparative studies are necessary to explore this further. Certainly, the effectiveness of the feedback strategies also depends on other factors to be investigated, such as the topic or the difficulty of the task and the individual prerequisites of the learners. A final decision between the two strategies may not be required by task creators, as strengths of the two feedback strategies may differ in cognitive, motivational and metacognitive areas.

References

Alarfaj, M., & Sangwin, C. (2022). Updating STACK Potential Response Trees Based on Separated Concerns. *International Journal of Emerging Technologies in Learning (iJET)*, *17*(23), 94–102. https://doi.org/10.3991/ijet.v17i23.35929

Altieri, M., Horst, J., Kallweit, M., Landenfeld, K., & Persike, M. (2020). *Multi-step procedures in STACK tasks with adaptive flow control*. https://doi.org/10.5281/zenodo.3944786

Bandura, A. (1997). *Self efficacy: The exercise of control*. W. H. Freeman.

Erickson, J. A., Botelho, A. F., McAteer, S., Varatharaj, A., & Heffernan, N. T. (2020). The automated grading of student open responses in mathematics. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 615–624. https://doi.org/10.1145/3375462.3375523

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*(1), 121–162. https://doi.org/10.1007/s10648-018-9446-6

Große, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction*, *17*(6), 612–634. https://doi.org/10.1016/j.learninstruc.2007.09.008

Hattie, J. (2010). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement* (Reprinted). Routledge.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Knaut, J., Altieri, M., Bach, S., Strobl, I., & Dechant, K. (2022). A Theory-Based Approach of Feedback in STACK-Based Moodle Quizzes Taking into Account Self-Regulation and Different Proficiency of Learners. *International Journal of Emerging Technologies in Learning (iJET)*, *17*(23), 38–55. https://doi.org/10.3991/ijet.v17i23.36425

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, *1*(4), 279–308. https://doi.org/10.1007/BF01320096

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In van Merrienboer, J. M. Spector, M. D. Merrill, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 125–144). Lawrence Erlbaum.

Narciss, S. (2012). Feedback Strategies. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 1289–1293). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_283

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multi-media learning. In H. M. Niegemann, D. Leutner, & R. Brünken (Eds.), *Instructional Design for Multimedia learning* (pp. 181–195). Waxmann.

Pals, F. F. B., Tolboom, J. L. J., & Suhre, C. J. M. (2023). Development of a formative assessment instrument to determine students' need for corrective actions in physics: Identifying students' functional level of understanding. *Thinking Skills and Creativity*, *50*, 101387. https://doi.org/10.1016/j.tsc.2023.101387

Ramdass, D., & Zimmerman, B. J. (2008). Effects of Self-Correction Strategy Training on Middle School Students' Self-Efficacy, Self-Evaluation, and Mathematics Division Learning. *Journal of Advanced Academics*, *20*(1), 18–41. https://doi.org/10.4219/jaa-2008-869

Renkl, A. (2014). The Worked Examples Principle in Multimedia Learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (2nd ed., pp. 391–412). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.020

Sangwin, C. (2023). Running an Online Mathematics Examination with STACK. *International Journal of Emerging Technologies in Learning (iJET)*, *18*(03), 192–200. https://doi.org/10.3991/ijet.v18i03.35789

Tärning, B., Lee, Y. J., Andersson, R., Månsson, K., Gulz, A., & Haake, M. (2020). Assessing the black box of feedback neglect in a digital educational game for elementary school. *Journal of the Learning Sciences*, *29*(4–5), 511–549. https://doi.org/10.1080/10508406.2020.1770092

Students’ instrumentalizations of hints and automated feedback in their task solution process when learning mathematics with a digital curriculum resource

Sebastian Rezat

Paderborn University, Institute of Mathematics, Germany; srezat@math.upb.de

Automated feedback is a characteristic feature of digital curriculum resources. Recently, there has been a growing interest in students’ perspectives on feedback. Feedback is increasingly regarded as a dialogic process in which learners make sense of information from varied sources and use it to enhance the quality of their work or their learning strategies. This study aims to further contribute to the body of research on the learners’ perspectives on feedback, by investigating for what purpose students make use of hints and automated feedback when learning mathematics with a digital curriculum resource. Students’ use of hints and automated feedback is analyzed through the lens of instrumental genesis. Results from a qualitative study with eight 8th-grade students show which type of hint or feedback is used at what phase for what purpose in the process of solving tasks from a widely available online curriculum resource.

Keywords: Automated feedback, hints, user study, students, digital curriculum resources.

Introduction

Automated feedback is a characteristic feature of digital curriculum resources (Choppin et al., 2014; Rezat, 2020). It aims to support students individually in their learning processes. The important role of feedback in learning is widely acknowledged (Hattie & Timperley, 2007). For a long time, feedback was regarded as a unidirectional process in which learners are viewed as receivers of information from an external source that they use to enhance their learning. Consequently, most research on feedback focuses on variables of the feedback message, such as the contents or the timing of when it is provided. Only in the past years, there has been a growing interest in students’ perspectives on feedback (Esterhazy & Damşa, 2019; Molloy & Boud, 2014; Olsson, 2018). Feedback is increasingly regarded as a dialogic process in which learners make sense of information from varied sources and use it to enhance the quality of their work or their learning strategies (Carless, 2015). Consequently, the meaning of feedback is not only determined by the feedback message, but by both, the agent and the user (Esterhazy & Damşa, 2019). Accordingly, there is a growing interest in better understanding how students seek, interpret, and use information related to their learning and how programs are designed to foster this (Molloy & Boud, 2014).

Elsewhere, I have shown how feedback can afford or constrain students’ conceptual development (Rezat, 2021) and that the interpretation of signs related to feedback on the artifact level imposes additional challenges on students (Rezat et al., 2021). This study aims to further contribute to the body of research on the learners’ perspective on feedback, by investigating the research question: For what purpose do students make use of hints and automated feedback when learning mathematics with a digital curriculum resource? The focus here is on the purpose of using the supportive information provided by the digital curriculum resource.

Theoretical framework

The instrumental approach (Rabardel, 2002) theorizes cognitive aspects of human interactions with digital artifacts and has proven useful in understanding students' learning of mathematics with Computer-Algebra and Dynamic Geometry Systems. In this paper, hints and feedback are considered artifacts developed to support students individually in their learning of mathematics. To develop a detailed account of how hints and feedback function in the learning process, students' use of these artifacts is analyzed through the lens of the instrumental approach.

Instrumental approach

According to Rabardel (2002) an artifact is transformed into an instrument in use. An instrument is a psychological entity that consists of an artifact component and a scheme component. In using the artifact, the subject attributes functions to the artifact and develops or adjusts utilization schemes that are shaped by both, the artifact and the subject. Attributing functions and the development of utilization schemes are two opposite but intertwined processes, which Rabardel refers to as "instrumentalization" and "instrumentation". In this paper, the focus is on the different functions that students attribute to hints and feedback in their learning process while solving tasks in an online learning platform and thus on their *instrumentalization*. Rabardel (2002, p. 106) defines *instrumentalization* as a "process in which the subject enriches the artifact's properties". Although this process is grounded in the artifact's intrinsic characteristics and properties, it is mainly linked to the subject's goals and conditions for action in a situation.

Hints and feedback

Feedback is widely defined as "information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (Hattie & Timperley, 2007, p. 81). Therefore, it only relates to information provided to students after they have solved a task and entered a solution into the system. However, many digital curriculum resources also offer information that learners can access on their way to their first solution before they have entered it into the system. This information is widely referred to as hints, cues, prompts, or tips. Referring to conceptualizations of feedback as a dialogic process in which learners make sense of information from varied sources and use it to enhance the quality of their work or their learning strategies (Carless, 2015) it makes sense to include this information in the analysis. Therefore, this paper analyzes students' use of supportive information provided by a widely used online curriculum resource that is offered in addition to the task itself before or after students have entered a solution to a task into the system. Consequently, clear conceptualizations of both, hints and feedback are necessary.

Whether a given task is a routine task or a problem does not only depend on task features but also students' knowledge and abilities. If students do not know how to find the solution to a given routine task immediately, the task becomes a problem for students. Zech (2002) suggests a taxonomy of five levels of hints that might support students in their problem-solving process: 1) motivational hints that motivate learners to continue the problem-solving process, 2) feedback that informs learners about the correctness of the selected solution strategy or achieved intermediate steps towards the solution, 3) general strategic hints that suggest a general problem-solving strategy to learners, 4) content specific strategic hints that provide learners with information about the solution strategy for the problem at hand, 5) content related hints provide learners with particular content that is relevant for solving the problem. A comparison between these types of hints and the different types of feedback in the next paragraph will show unmistakable overlaps in their content. Additionally, the digital curriculum resource used in this study sometimes offers the same information before and after entering the solution. Therefore, depending on when the information is presented it would be either considered a hint or elaborated feedback. Consequently, I use the same terminology to distinguish the different types of hints and feedback.

To differentiate different types of feedback offered by digital curriculum resources, the study presented in this paper refers to the classification of feedback according to Shute (2008). Shute distinguishes different types of feedback according to their complexity. The following types are relevant:

1. Knowledge of results feedback (*KR*) informs the learner about the correctness of an answer.
2. Knowledge of correct response (*KCR*) feedback informs the learner about the correct response.
3. Repeat-until-correct (*RUC*) feedback informs the learner about an incorrect response and offers the possibility of a new try to answer the task.
4. Location of mistakes (*LOM*) feedback informs the learner about the location of an error in the solution without giving the correct response.
5. Elaborated feedback (*EF*) offers further information regarding the solution of the task or the solution of the learner.

For the last type (EF), the literature distinguishes many different subtypes. For the study presented in this paper, the following types of *EF* are relevant that Shute (2008, p. 160) subsumes under "topic contingent" and "hints/clues/prompts":

- knowledge about concepts (*kac*),
- knowledge or strategic information on how to proceed (*kohp*),
- a worked example or demonstration (*we*)

Types of hints and feedback in the used digital curriculum resource

As apparent from the theoretical framework, hints and feedback are characterized differently and use different terminology. However, the contents of the messages seem to be equivalent in many cases. Additionally, *bettermarks* shows the same message sometimes as a hint and sometimes as feedback. To develop a clear terminology to denote the information provided by hints or feedback messages, the types of hints according to Zech (2002), types of feedback according to Shute (2008), and the features of the used digital curriculum resource (DCR) that contain these types of hints and feedback are juxtaposed in Table 1. This juxtaposition reveals the overlaps in terms of the content of the feedback message. However, the scope of some of the types is different. On the one hand, the differentiation between general and content-specific strategic hints is not mirrored in different types of feedback, on the other hand, both, elaborated feedback presenting knowledge on how to proceed, or a worked example can be considered content-specific strategic hints. Comparing the types of hints and feedback with the related features in the curriculum resource shows that a single feature may present a variety of different types of hints or feedback. Especially the feature "Tip" may provide knowledge about concepts, knowledge on how to proceed, or a worked example.

Table 1: Juxtaposition of hints, types of feedback, and related features in the used DCR

Type of hint	Type of feedback	Appearance in the DCR
1) motivational hints	-	-
2) hints that provide learners with feedback about the correctness of the selected solution strategy or achieved intermediate steps toward the solution	KR-feedback	KR-feedback after a solution was entered
3) general strategic hints	EF (kohp)	-
4) content specific strategic hints	EF (kohp)	Feature called “Tip” Sometimes shown automatically after entering a wrong solution
	EF (we)	Feature called “Lookup” Feature “Example” sometimes appearing after a wrong solution
5) content related hints	EF (kac)	Feature called “Tip” Linked technical terms in the task

Methodology

This study aims to analyze students' use of hints and feedback from digital curriculum resources in an ecologically valid setting. Therefore, the widely used resource *bettermarks* (www.bettermarks.com) was used. *Bettermarks* offers a wide range of different types of hints and feedback while solving a task. It is licensed in several federal states in Germany by the federal ministries of education and thus offered for free to schools. In this study, eight students in eighth grade were working on a unit on percentages. The unit comprised a set of 17 tasks and problems that students worked through at the end of the unit on percentages as a preparation for a test. The tasks were carefully selected to comprise different kinds of problems covering all the content that was relevant for the test and offering to students possibly all the different kinds of hints and feedback available in *bettermarks* (at that time). The eight students worked on the set of tasks at home in their familiar setting in a video conference with the interviewer on a shared screen. The students were used to working with *bettermarks* at home. Thus, the situation of using *bettermarks* was kept as natural as possible. The only difference was the presence of the interviewer in the video conference.

The recordings of the video conferences provide the data for this study. The videos were analyzed using the qualitative data analysis software MAXQDA. In the first step, each video was coded for the different types of hints and feedback used by the students. In the second step, each episode in which the students used a hint or received feedback was analyzed in terms of the purpose that the students associated with its use. This was done based on the constant comparative method (Corbin & Strauss,

2015) until different *instrumentalizations* could be delineated and defined. As instrumentalizations refer to functions that students attribute to the hints and feedback by their goals and conditions for action in a situation, this was achieved by inferring students' motivations or reasons and their goals for using a hint or feedback from the data. These partly depend on the phase in the solution of the problem, in which students make use of the hint or feedback.

Results

Table 2 provides an overview of the results of the analysis of students' instrumentalizations of the different types of hints and feedback that are offered by *bettermarks*. The left column of Table 2 is organized in chronological order and describes the different phases that students must go through when solving a task from *bettermarks*. These different phases characterize the situations in which hints and feedback are used. As described in the theoretical framework, these influence students' instrumentalization of hints and feedback. The second column contains the different types of hints or feedback that are offered by *bettermarks* in the different phases and were used by the students. In phases 1–3, the types characterize the contents of hints. Starting in phase 4, the types relate to feedback. The third column shows students' *instrumentalizations* of these types of hints and feedback in the respective phase.

Table 2 may be read in the following way: While reading the task (phase 1) *bettermarks* offers hints of *kac*-type. Students instrumentalize these hints to enhance their understanding of technical terms that appear in the tasks. In phase 2, when students aim to find the solution to the task, they have access to three different kinds of hints: They can open *kac*-type or *kohp*-type hints or they can ask for the complete solution of the task (KCR). *Kac*-type and *kohp*-type hints are instrumentalized in two different ways: Students either use them to get support in finding the solution or to resolve uncertainties about the expected input format when entering the solution into the system. KCR is used to understand the expected solution if students do not develop any solution on their own. Starting in phase 4, columns 2 and 3 are divided as there are two possibilities for feedback depending on the correctness of the entered solution. The same applies to the fields in phase 7. Fields shaded in grey denote that no more feedback or hints are accessible at these phases.

Table 2: Cumulated results of students' instrumentalizations of different types of hints and feedback in the different phases of the task solution process

Phase in the solution process	**Type of hint / feedback**	**Instrumentalization**
1. Reading task	kac	Enhancing understanding of technical terms in the task
2. Finding solution	kac kohp	Getting support to find the solution
		Resolving uncertainties about the expected input format
	KCR	Understanding the expected solution

<table>
<tr><td><b>3. Entering solution</b></td><td colspan="3">kac
kohp</td><td colspan="3">Confirming that the solution (procedure) is correct before entering</td></tr>
<tr><td rowspan="3"><b>4. Evaluation of the entered solution</b></td><td rowspan="2">KR feedback: correct</td><td colspan="2" rowspan="3">KR feedback: incorrect</td><td>Reassurance of own solution</td><td colspan="2">Removing uncertainties about which solution from two alternatives is the correct one</td></tr>
<tr><td colspan="3">Resolving uncertainties about the expected input format</td></tr>
<tr><td>KCR</td><td>Checking own / alternative solution</td><td colspan="2"></td></tr>
<tr><td rowspan="2"><b>5. Rethinking solution</b></td><td rowspan="5"></td><td colspan="2">KCR</td><td rowspan="5"></td><td colspan="2">Understanding the expected solution
Understanding own mistakes</td></tr>
<tr><td colspan="2" rowspan="2"><i>Hints as in phases 2 & 3</i></td><td colspan="2" rowspan="2"><i>Instrumentalizations as in phases 2 & 3</i></td></tr>
<tr><td><b>6. Entering adjusted solution</b></td></tr>
<tr><td rowspan="2"><b>7. Evaluation of solution</b></td><td>KR
correct</td><td>KR
incorrect</td><td colspan="2"><i>No particular instrumentalization observed</i></td></tr>
<tr><td></td><td>KCR</td><td></td><td>Understanding the expected solution
Understanding own mistakes</td></tr>
</table>

Discussion

Many of the *instrumentalizations* of hints and feedback can be expected and seem to fit the intended purpose of the type of hint or feedback. This is for example the case for the following hint/feedback/instrumentalization pairs: (*kac*&*kohp*/getting support to find the solution) or (*KCR*/understanding the expected solution). However, some instrumentalizations are particularly interesting as they indicate difficulties that students have with solving tasks from a DCR. This is especially the case for *instrumentalizations* related to issues with the input format. In these cases, students are not sure, what kind of input or input format (e.g., fraction or decimal, exact or rounded decimal) is expected by the system. They either instrumentalize hints to resolve these uncertainties before entering the solution or the *KR* feedback helps to resolve the uncertainties after entering the solution. Another unexpected instrumentalization is that some students who have found the solution to a task instrumentalize use hints before entering the answer to ensure that their answer is correct instead of simply using the *KR*-feedback for this purpose. However, these utilizations are closely related to the constraints of the DCR. They are mostly not of mathematical relevance. For example, if a number is written as a fraction or a decimal is equivalent from a mathematical point of view, but the DCR only accepts one input as correct.

On the one hand, the results show that different kinds of hints and feedback are instrumentalized for the same purpose. For example, students use *kac* or *kohp* hints or *KR*-feedback to resolve uncertainties about the expected input format. However, as the results give a cumulated overview of the observed instrumentalizations of all participating students, nothing can be said about whether this is the case for one particular student or if this is an observation that only appears between different students. A deeper, case-sensitive analysis is necessary to reveal if a particular student shows a definite *instrumentalization* of a particular type of hint or feedback for a specific purpose. If this is not the case, i.e. if one student uses different types of hints and feedback for the same purpose it may be an indication of an incomplete instrumental genesis. However, it may also be an issue with the system, as it is not always clear what kind of hint is provided by *bettermarks* when students look for support while solving a task especially when they use the feature “Tip”. Furthermore, the analysis only focused on what type of hint or feedback was used for a particular purpose. It was not analyzed if the hint or feedback was actually supportive in the sense that it either helped to solve the problem or if students thought that the information was helpful. This would also be a matter of deeper analysis.

On the other hand, the results show that one type of feedback is instrumentalized for different purposes. For example, *KCR*-feedback is instrumentalized for two different purposes: First, for understanding the expected solution before or after entering a solution when students do not have a clue of how to solve the tasks. Second, students also instrumentalize *KCR*-feedback after entering the correct answer to compare their solution with the provided one to check their solution procedure or to see an alternative solution.

In summary, the results show that students instrumentalize the different types of hints and automated feedback offered by the system when solving tasks from a DCR for different purposes and that one type of hint or feedback may be instrumentalized at different phases of the solution process for different purposes. Consequently, these results underline the starting point of the study, namely that “the meaning of feedback is not only determined by the feedback message, but by both, the agent and the user” (Esterhazy & Damşa, 2019). A deeper analysis may reveal how students’ instrumentalizations of hints and feedback contribute to a successful solution of the task and their learning of mathematics. These insights could be helpful for the design of hints and automated feedback in DCR as they show students’ difficulties and needs during their individual learning processes.

References

Carless, D. (2015). *Excellence in university assessment: Learning from award-winning practice*. Routledge.

Choppin, J., Carson, C., Borys, Z., Cerosaletti, C., & Gillis, R. (2014). A typology for analyzing digital curricula in mathematics education. *International Journal of Education in Mathematics, Science and Technology*, *2*(1), 11–25.

Corbin, J., & Strauss, A. (2015). *Basics of qualitative research: Techniques and procedures for developing Grounded Theory* (4th ed.). Sage.

Esterhazy, R., & Damşa, C. (2019). Unpacking the feedback process: An analysis of undergraduate students' interactional meaning-making of feedback comments. *Studies in Higher Education*, *44*(2), 260–274. https://doi.org/10.1080/03075079.2017.1359249

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Molloy, E. K., & Boud, D. (2014). Feedback models for learning, teaching and performance. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 413–424). Springer. https://doi.org/10.1007/978-1-4614-3185-5_33

Olsson, J. (2018). The contribution of reasoning to the utilization of feedback from software when solving mathematical problems. *International Journal of Science and Mathematics Education*, *16*(4), 715–735. https://doi.org/10.1007/s10763-016-9795-x

Rabardel, P. (2002). *People and technology: A cognitive approach to contemporary instruments* https://hal.archives-ouvertes.fr/hal-01020705/document

Rezat, S. (2020). Mathematiklernen mit digitalen Schulbüchern im Spannungsfeld zwischen Individualisierung und Kooperation. In D. M. Meister & I. Mindt (Eds.), *Mobile Medien im Schulkontext* (pp. 199–213). Springer Fachmedien. https://doi.org/10.1007/978-3-658-29039-9_10

Rezat, S. (2021). How automated feedback from a digital mathematics textbook affects primary students' conceptual development: Two case studies. *ZDM – Mathematics Education*, *53*, 1433–1445. https://doi.org/10.1007/s11858-021-01263-0

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Zech, F. (2002). *Grundkurs Mathematikdidaktik. Theoretische und praktische Anleitungen für das Lehren und Lernen von Mathematik* (10 ed.). Beltz.

Professionalising teachers with a digital formative assessment tool: a case study of the SMART tests

Fabian Rösken and Bärbel Barzel

University of Duisburg-Essen, Germany; fabian.roesken@uni-due.de

SMART, a digital formative assessment tool for mathematics education, aids teachers by offering short online diagnostic checks and detailed analyses of response patterns. It identifies student misconceptions and provides teachers with valuable insights on students' level of understanding. This should have the potential to not only influence teaching and student understanding but also the teachers' professionalisation. This paper explores the potential impact of using SMART on five mathematics teachers' PCK in elementary algebra. Initial findings indicate the tool's effectiveness in enhancing teachers' knowledge of student thinking and typical mistakes.

Keywords: Technology-supported formative assessment, PCK, teacher professionalisation

Theoretical background

Formative assessment has been proven to be a valuable approach for students to benefit in their learning. Andrade & Heritage (2018) concretise that a successful implementation of formative assessment consists of a diagnosis of students' performance and progress, and an adjustment of teaching to the individual needs, for example by reacting to diagnosed misconceptions. Busch et al. (2015) have investigated teachers' formative assessment practices and found out, that content specific pedagogical content knowledge (PCK) has a high impact on diagnostic practices. They observed a shift from superficial diagnoses (e.g. just correcting mistakes) to a deeper analysis as soon as aspects of content specific PCK are shown. The PCK of teachers is often divided into different equally meaningful aspects, e.g. in the COACTIV study it was defined as *knowledge of explaining and representation*, *knowledge of student thinking and typical mistakes*, and *knowledge of the potential of mathematical tasks.* This framework has demonstrated predictive validity in enhancing instructional quality and fostering student learning gains across various studies (Krauss et al., 2020).

Technology has the potential to support teachers in formative assessment processes while also developing the needed PCK of teachers (Stacey & Wiliam, 2013). One example of technology-based formative assessment is the SMART system, which provides a fast and in-depth diagnosis by specifically designed diagnostic items and their deep analysis (Stacey et al., 2018). After students complete a 5- to 10-minute test, only the teachers receive an automated diagnosis for each student in the form of comprehension levels and misconceptions. In addition, further explanations and teaching suggestions are provided, which include tips for teaching, general advice on rituals, attitudes, methods and desirable concepts, as well as concrete tasks that can be used directly to support the individual students (Price et al., 2013).

In this study we used a test from the field of algebra. The test *Meaning of Letters* examines whether students consistently interpret variables in such a way that they stand for numerical values and not as an abbreviation for objects occurring in the context. The test includes six multiple-choice items whose incorrect answer options reflect typical misinterpretations. For example, when the question is *"Biros are sold in packs of 3. Sam bought p packs and got b biros altogether. Choose the correct equation"*, the answer option *“3b = p”* can be understood as *“There are 3 biros in one pack”*, where the variables are interpreted as abbreviations for biros and packs respectively. This is known as the *letter-as-object* (LO) misconception (Stacey et al., 2018). This misconception is especially important since it can linger with students as they navigate their way through further algebra. Students who struggle to accurately formulate equations and expressions to represent real-world scenarios may miss out on harnessing the problem-solving potential of algebra. Given the accessibility of digital technology capable of solving equations, it becomes increasingly crucial for students to master the skill of constructing equations and accurately interpreting equations created by others, rather than focusing solely on solving equations manually (Arcavi et al., 2017). The SMART test compares the answer to the Biro item with the responses to similar items and searches for specific answer patterns to reveal if a student has this misconception.

The automatic diagnosis is displayed to the teachers in the form of three comprehension levels, based on the frequency on which this misconception is revealed. One additional misconception that the students may hold could also be flagged. The solution-as-coefficient (SAC) misconception is a subtype of the LO misconception, where a possible solution for an equation is already found and placed in front of the variables as coefficients. It is particularly important that teachers are able to recognise these misconceptions and know how to address them. Teachers who recognise that existing misconceptions may impede algebra learning can support their students by openly addressing the distinctions between coding, other notation systems, and the realm of algebra. Therefore the PCK aspect *knowledge of student thinking and typical mistakes* is particularly important for successful algebra teaching.

Research design

The nature of SMART tests suggests that teachers increase their *knowledge of student thinking and typical mistakes* while using SMART, because they engage with their own students’ misconceptions and possible ways to overcome them. The aim of this study was to find out, to what extent a development of content specific PCK among teachers can be seen after using the SMART tests. To answer this question, five teachers of grade 7 participated in a pre-intervention-post-test design. The *knowledge of student thinking and typical mistakes* was surveyed in a pre-test with a competency test. As an intervention, they used the SMART test *Meaning of Letters* with their students and, based on the results, taught the topic of variables and their meaning for about 4 weeks. They then used a second version of this SMART test with their students to track student development. As a post-test, the teachers were surveyed with a second version of the competency test.

To test the PCK competency, an existing test by Busch et al. (2015) was adapted for algebra and understanding of variables. The analysis of the COACTIV study revealed that written tests were highly predictively valid for individual learning support (Krauss et al., 2020). In the competency test, four example student solutions were generated for teachers to assess. The examples represent typical student solutions to tasks on the understanding of variables and the concept of terms and equations and contain frequent misconceptions. Table 1 shows an example of one of the four examples in which the LO misconception is present. With the relevant *knowledge of student thinking and typical mistakes*, it can be recognised that this student does not interpret the variable as a placeholder for a numerical value, but rather as an abbreviation for an object (*m* stands for marshmallows).

Additionally, the students' solution includes the correct answer as the equation (SAC) and is likely to be read as a solution sentence ("*4 marshmallows and 5 caramels together cost 80 cents*").

Table 1: Example student solution "Candy shop" from the pre-test

As a task to formulate linear equations, Anthony received the following task:
Catherine went into a candy shop: "I have bought marshmallows and caramels and paid a total of 80 cents. The marshmallows cost 10 cents each and the caramels 8 cents." Formulate an equation, that describes this situation.
Anthony wrote: $4m + 5c = 80$

The teachers were asked to formulate their own diagnosis as well as approaches for spontaneous and further supportive teaching practice. The teachers' answers were analysed in how far they could describe the possible student thinking and if they – implicitly or explicitly – refer to the underlying misconceptions the student may hold. The analyses of the pre- and post-tests were compared and examined for a development of PCK. The teachers' statements were qualitatively analysed by the first author and interrated by at least one other person.

Results

After examining the cases in how far they show *knowledge of student thinking and typical mistakes*, we were able to contrast three different ways of development: (1) two teachers showing implicit knowledge in the pre-test and becoming more explicit in the post-test, (2) two teachers showing no corresponding knowledge in the pre-test and showing implicit knowledge in the post-test, and (3) one teacher who constantly shows a lot of explicit knowledge.

As stated above, two teachers exhibited an implicit *knowledge of student thinking and typical mistakes*, primarily focusing on describing and evaluating the student's solution. For instance, they acknowledged the student's grasp of the situation but noted the inclusion of an unnecessary solution in the equation. Implicitly addressing the SAC misconception, the teachers lacked further explanation, failing to identify the underlying LO misconception. This gap was evident in the superficial and process-oriented supportive practices they expressed, in which they emphasised to highlight key information of the text and practice general strategies for text tasks. In the post-test, the teachers can identify slightly more precise that the student shows difficulties with the meaning of the variables. This becomes evident when they express possible supportive approaches in the post-test, because they focus on the meaning of the variable and suggest to "*write down what m and c mean*" and "*use simple number examples to check the solution*".

A similar development can be seen for two of the teachers who not only remained on a superficial focus in the pre-test ("*The equation does not make sense*"), but sometimes even diagnosed incorrectly and did not perceive the student's error at all. Accordingly, the supportive teaching practices remained predominantly at a general, motivational level ("*He should read the assignment again carefully*") or at a process-oriented level, vaguely suggesting that they would try to solve the equation together with the student. In the post-tests, these teachers showed a greater understanding of the student's difficulties, and the approaches show a more concrete connection to the task and the student's errors.

For example, they suggest that the student should write the equation again without using the initial letters. This hints that these teachers show an implicit understanding of the LO misconception.

In one case, a teacher showed a consistently high *knowledge of student thinking and typical mistakes* in pre- and post-test ("*For Anthony, m seems to stand for the object marshmallows, and he is probably thinking of a spoken form of the equation: '4 marshmallows and 5 caramels cost 80 cents.' He apparently already has a solution in his head.*").

Discussion

The preliminary findings on SMART's impact on teachers' PCK are promising, as progress was seen for all teachers regarding the PCK aspect *knowledge of student thinking and typical mistakes*. While initial diagnoses had few variations, all teachers showed more explicit knowledge of underlying misconceptions to students' errors in the post-test. Those with implicitly recognisable knowledge in the pre-test became more explicit, utilising content specific terms. Teachers that struggled to explain student thinking in the pre-test could show implicit knowledge of the LO misconception in the post-test. Teachers with a lot of corresponding knowledge maintained this precision. These results align with the expectation that SMART increases teachers' *knowledge of student thinking and typical mistakes* and strengthens them to avoid an exclusive procedural knowledge focus. The presented cases are part of a broader study with over 40 teachers receiving various distinct interventions.

References

Andrade, H. L., & Heritage, M. (2018). *Using formative assessment to enhance learning, achievement, and academic self-regulation.* Routledge

Arcavi, A., Drijvers, P. & Stacey, K. (2017). *The learning and teaching of algebra*. Routledge.

Busch, J., Barzel, B. & Leuders, T. (2015). Die Entwicklung eines Instruments zur kategorialen Beurteilung der Entwicklung diagnostischer Kompetenzen von Lehrkräften im Bereich Funktionen. *Journal für Mathematik-Didaktik*, 36(2), 315–338.

Krauss, S., Bruckmaier, G., Lindl, A., Hilbert, A., Binder, K., Steib, N. & Blum, W. (2020). Competence as a continuum in the COACTIV study: the "cascade model". *ZDM Mathematics Education,* 52, 311–327.

Price, B., Stacey, K., Steinle, V., & Gvozdenko, E. (2013). SMART online assessments for teaching mathematics. *Mathematics Teaching*, 235(4), 10–15.

Stacey, K., Steinle, V., Price, B., & Gvozdenko, E. (2018). Specific mathematics assessments that reveal thinking: An online tool to build teachers' diagnostic competence and support teaching. In T. Leuders, J. Leuders, K. Philipp, & T. Dörfler (Eds.), *Diagnostic competence of mathematics teachers – Unpacking a complex construct in teacher education and teacher practice* (pp. 241–263). Springer

Stacey, K., & Wiliam, D. (2013). Technology and assessment in mathematics. In M. A. K. Clements, A. J. Bishop, C. Keitel, J. Kilpatrick & F. K. S. Leung (Eds.), *Third international handbook of mathematics education* (pp. 721–751). Springer.

Formative feedback in problem-solving lessons in German primary schools

Yasmin Theile[1] and Benjamin Rott[2]

[1,2] University of Cologne, Germany; yasmin.theile@uni-koeln.de

In this article, we present initial results of the research on formative feedback, which is used by teachers in problem-solving orientated lessons in German primary schools. The analyses of five lessons on problem solving from German primary schools revealed 15 different forms of formative feedback from different levels (Hattie & Timperley, 2007) which will be exemplified. Most of the feedback forms identified relate to the task, results or process. Formative Feedback regarding the level of self-regulation could not be identified.

Keywords: Problem solving, primary education, formative feedback.

Introduction

We all need people who will give us feedback. That´s how we improve.

Bill Gates

Problem solving (PS) is a one of the main activities of mathematical work and therefore an important skill to be learned in school mathematics. It is uncontroversial that students should learn to solve problems from primary school onwards and to reflect on and, if necessary, adapt their approaches to PS. However, studies have shown that primary school students in particular have problems in performing these skills (Heinrich et al., 2014). As a result, there is an increased need for teacher support to improve students' PS skills. The opening quote supports the importance of feedback for improvement – feedback can therefore be seen as a starting point for promoting problem-based teaching. As the particular complexity of everyday teaching makes diagnosis and specific support for students difficult (Heinrichs & Kaiser, 2018), this study focuses on formative feedback from teachers.

Theoretical Background

Feedback is one of the key strategies of formative assessment (Black & William, 2009), which can be defined as a *practice* that is

> formative to the extent that evidence about student achievement is elicited, interpreted, an used by teachers, learner, or their peer, to make decisions about the next step in instruction that are likely to be better, or better founded, than the decision they would have taken in the absence of the evidence that was elicited (Black & Wiliam, 2009, p. 9).

Feedback can generally be "conceptualized as information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (Hattie & Timperley, 2007, p. 81). In the field of educational research, a distinction is also made between formative and summative feedback. Formative feedback aims to record the current learning status of the students and, based on this, to adapt the teaching and control the learning process of the students (Black & Wiliam, 2009; Polly et al., 2017). Hattie and Timperley (2007) also distinguish between four different levels of formative feedback. Feedback at the personal level is aimed directly at students and mainly includes praise or motivation and is often unrelated to the task or performance of the student (Hattie, 2012; Hattie & Timperley, 2007). Feedback at the task and product level includes all feedback and information about how well a task has been mastered or performed (Hattie & Timperley, 2007). It is also often called corrective feedback and is the most common form of feedback given in a classroom setting (Hattie, 2012). Feedback at the process level refers to the processes on which the execution of a task is based. It "can lead to providing alternative processing, reducing cognitive load, [or] helping to develop learning strategies and error detection" (Hattie, 2012, p. 119). The highest form of feedback takes place at the level of self-regulation. Feedback at this level refers to the students' ability to monitor their own learning processes (Hattie, 2012). Although the influence of feedback on learning is undisputed, there are so far – especially in relation to problem-based teaching – only a few empirical findings on the implementation of formative feedback in mathematics education (e.g. Green, 2023). Since studies have shown that the positive effects of formative feedback on learning could depend on the subject and specific implementation (McLaughlin & Yan, 2017), studies are needed specifically on problem-based mathematics teaching.

Research question and study design

To be able to make statements about the effectiveness of specific forms of feedback in problem-based teaching, it is useful to identify the forms of feedback that teachers use in everyday teaching. Amongst this background, the aim of the study is to investigate how teachers support their students' solution processes in problem-based lessons. This article will focus on the question which forms of formative feedback can be identified among primary school teachers in problem-based lessons. Tor this study, five lessons were videotaped in different primary schools in which the predefined aim of the lessons was to work on a problem-based task. For this purpose, the teachers were given a catalogue of problems to choose from. Three teachers taught a 2nd-grade class, the other two 4th-grade classes; more lessons are currently recorded and evaluated. Each lesson (45 min) was videotaped using two cameras to cover the classroom and a microphone or GoPro for the teacher. No further specifications were given to the teachers in advance.

The analysis is based on Mayring's (2000) qualitative-content-analysis methodology, allowing for an deductive-inductive category development. To develop a category system, the levels of feedback (Hattie & Timperley, 2007) were used as superordinate categories to provide a basic structure. The various forms of feedback that could be identified in the analysis were then described in detail, categorized by similar characteristics and subordinated to the levels of feedback. The previous analyses were performed by one coder.

Results

We identified a total of 176 situations in which teachers gave feedback to students on their work. These could be summarized into 15 feedback categories, which in turn could be categorized into three of the four feedback levels. Only three, because feedback at the level of self-regulation has not been identified in the so-far recorded data. One category, the feedback form *R_1: Reception signal,* could not be assigned to any of the four levels and was therefore coded as Unspecific reactions (Rx). Typical reception signals that could be observed were *hm, mhm, aha* or emotional expressions such as laughter as reactions. The category *T_1: Correctness of Result* – categorized at the task and product level (*T_x*) – is coded when the teacher explains to the students that a (partial) result is correct, incorrect or incomplete without providing further information. As this category could be identified in the lesson of each teacher, it will be shown as an example below. Mrs. P's interaction with a student illustrates this. The students must find the number of squares that can be found on a chessboard (204), which is likely to be a problematic task for many primary school students, as they do not have a suitable solution scheme is not immediately obvious. The student approaches her teacher during the lesson and states:

157 Student Mrs. P. .. I think I'm done, I have 118.
158 Mrs. P.: No its more.
159 Student: What, great.

After the last sentence, the student returns to her seat. The teacher does not provide any further information, such as what has not yet been correct or forgotten by the student. Interactions that could be assigned to this category were identified 24 times.

A complete list of all categories that could be identified is given in figure 1. It also shows how often each category occurred in the analyses. The letter in the abbreviation stands for the level to which the categories have been assigned – personal level (S_x), task and product level (T_x) and process level (P_x). The complete coding agenda will be gladly provided on request.

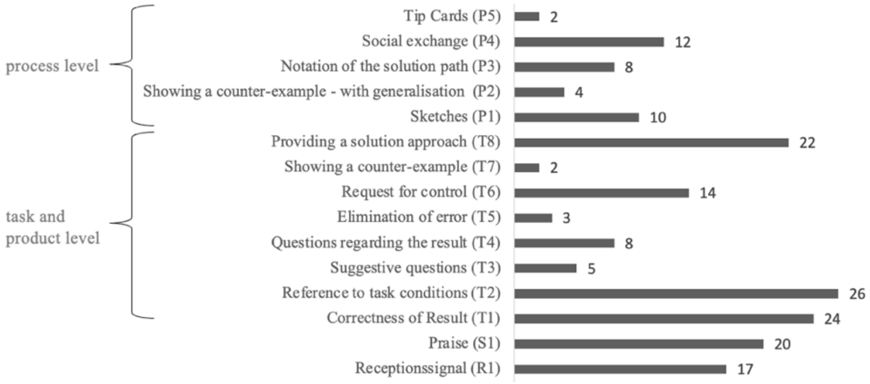


Figure 1: Categories of formative Feedback in problem-based lessons

It is noticeable that some forms of feedback such as T_2*: Reference to task conditions* and T_8*: Providing a solution approach* were identified very frequently and with every teacher, while others such as P_5*: Tip Cards* occurred only with individual teachers. However, it should be noted that these results may be due to the small amount of data in this study. The analysis of further lessons will provide more in-depth findings here. The results presented support the findings of Hattie (2012) that most feedback is given at the task and product level. Overall, it is noticeable that the teachers in this study most frequently chose forms of feedback that address the product level, although problem solving is a process-related skill and therefore the actual process should be the focus.

The study also revealed that certain forms of feedback were identified as recurring for a teacher. An exemplary case of this is Mrs. T. In the evaluation of her lesson, 40 feedback situations were identified in which she responded with one of three forms of feedback in around 64% of cases (T_1, T_2 and T_8). All other identified categories occurred only sporadically. While this phenomenon was recognizable among all teachers, the categories that occurred particularly frequently differed between the individual teachers. This will be analyzed in greater depth in further analyses.

Discussion

The aim of this study was to identify and analyze different forms of formative feedback which are used by German primary school teachers in lessons regarding problem solving. It was possible to identify 15 different categories, which can be assigned to three levels of feedback. The catalogue of criteria resulting from this study is currently being further tested for its applicability by expanding the data by analyzing additional lessons – including external analyzes by other coders. It should be noted that there may be further changes to the list of categories in the subsequent analyzes. In particular, categories at the level of self-regulation (Hattie & Timperley, 2007) that could not be identified in the current analyzes are conceivable here. Based on the findings so far, further analyses should also focus on exploring possible recurring preferred forms of feedback that occur across lessons for individual teachers. In the long term, it seems particularly interesting to examine what effect the forms of formative feedback used by the teachers have on students' problem-solving processes in order to allow specific conclusions for teaching practice.

References

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Green, J. (2023). Primary students' experiences of formative feedback in mathematics. *Education Inquiry*, *14*(3), 285–305. https://doi.org/10.1080/20004508.2021.1995140

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Heinrich, F., Jerke, A., & Schuck, L.-D. (2014). "Fehler" in Problembearbeitungsprozessen von Grundschulkindern. In J. Roth & J. Ames (Eds.), *Beiträge zum Mathematikunterricht 2014* (pp. 499–502). WTM.

Heinrichs, H., & Kaiser, G. (2018). Diagnostic Competence for Dealing with Students' Errors: Fostering Diagnostic Competence in Error Situations. In T. Leuders, J. Leuders, & K. Philipp (Eds.), *Diagnostic Competence of Mathematics Teachers. Unpacking a Complex Construct in Teacher Education and Teacher Practice* (pp. 79–94). Springer International.

McLaughlin, T., & Yan, Z. (2017). Diverse delivery methods and strong psychological benefits: A review of online formative assessment. *Journal of Computer Assisted Learning*, *33*(6), 562–574. https://doi.org/10.1111/jcal.12200

Polly, D., Wang, C., Martin, C., Lambert, R. G., Pugalee, D. K., & Middleton, C. W. (2017). The Influence of an Internet-Based Formative Assessment Tool on Primary Grades Students' Number Sense Achievement. *School Science and Mathematics*, *117*(3–4), 127–136. https://doi.org/10.1111/ssm.12214

Students’ feedback process using automated post-submission report in their modeling process

Muna Touma[1] and Shai Olsher[2]

[1]University of Haifa, Israel; munatouma89@gmail.com

[2]University of Haifa, Israel; olshers@edu.haifa.ac.il

In this study, we explored a pair of students’ interactions with the automated post-submission report to learn how they use this tool to deepen their own modeling process. The students worked on a modeling activity consisting of four example-eliciting tasks in which they were asked to construct position-over-time graphs. The findings show that by interacting with the post-submission report, new sequences of mathematical modeling competencies use were identified, which enables fostering the student’s interpreting and validating the mathematical results.

Keywords: Feedback process, mathematical modeling, modeling process.

Introduction and theoretical background

Feedback is an ongoing process in which learners make sense of information related to the task or process of learning, which is provided by an agent (e.g., teacher, peer, book) regarding aspects of one’s performance or understanding (Hattie & Timperley, 2007), to fill a gap between what is understood and what is aimed to be understood (Sadler, 1989). Indeed, studies have shown that the feedback given to students has a greater impact on their achievement than any other teaching strategy (Hattie & Timperley, 2007). In their study, Yerushalmy et al. (2023) discussed how artifacts, particularly personal feedback that are designed as tools by others, might become instruments for learning. They referred to personal feedback as a mirror that describes an instance by offering a task designer’s description of a learner’s example. This description includes the characterization of students’ examples in words.

Using technology in mathematics education affects the teaching and learning processes; specifically, it can be used to support the feedback process in many ways. Technology can be used to generate outputs such as numeric grades, written reports, and statistics for use by teachers (Sangwin & Köcher, 2016), students (Yerushalmy et al., 2023), or both (Abu Raya & Olsher, 2021; Sadler, 1989). Some technologies also have the feature of reporting to the teacher regarding the tools that students use when they solve the task (Abu Raya & Olsher, 2021), while others have the feature of analyzing the answers according to mathematical characteristics (Abu Raya & Olsher, 2021). In addition, the feedback process can be supported by the effective use of digital tools, such as simulations and interactive diagrams, which provide learners with immediate information that can help them solve problems and can positively affect various competences, such as understanding, validating, and interpreting, which are considered modeling sub-competences (Greefrath, 2011). Thus, technology can promote mathematical modeling, which is an important competency that students of all ages ought to acquire, owing to its role as a method for better understanding the world around us. Greefrath (2011) described two areas in which the modeling process occurs: the rest of the world and mathematics, and introduced a modeling cycle (MC) that represents key phases (in black font) and transitions between them (in red font) in the processing of reality-related problems. He suggests that digital tools may be useful at each step of this process (Figure 1). For example, for experimenting, by transforming a real situation into a geometrical model, with the help of dynamic geometry software.

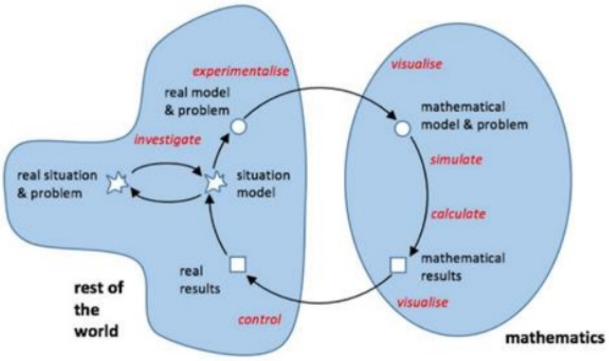


Figure 1: Modeling cycle with the added influence of digital tools (Greefrath, 2011)

Mathematical modeling literature highlights the need for more research to better understand how technology may assess or foster mathematical modeling competencies (MMC) (Cevikbas et al., 2022). Therefore, the purpose of this study is to gain insights into automated post-submission reports serving as a bridging tool between the literacy phenomenon and the given mathematical model. We asked: (1) How do students use post-submission reports while working on the modeling activity? (2) Do students' feedback processes using automated post-submission reports affect their MMC?

Methodology

Population

We present a case study involving a pair of 9th graders (14-year-old students), Ray and Lour (fictitious names), who completed a computer-based modeling activity with their teacher during mathematics lessons. The pair of students was chosen based on their educational level (medium-high), their good expressiveness, and their seriousness.

Tools

The research tools included a modeling activity designed on the STEP platform and class observations. The STEP platform was designed to support the assessment of various patterns of open-ended example-eliciting tasks. The students' submitted examples were analyzed online according to specified characteristics, producing a post-submission report for the students (Harel et al., 2019). These characteristics can provide information that goes beyond whether students' submissions are correct or not, can give students a sense of the breadth of their personal example space, and may be resources to create a shift in students' understanding (Harel et al., 2019).

The modeling activity "Cycling" used in this study was designed on the STEP platform according to specific design principles (for further details, see Touma and Olsher, 2022). It includes four example-eliciting tasks that are designed so that the first task is an introductory task that helps students familiarize themselves with the content and tools, and the advanced tasks offer opportunities for higher cognitive demands. Some of the tasks include digital simulations and tools that allow interaction with the various components of the model, offloading part of the student's mathematical work to these tools (Touma and Olsher, 2022). This activity deals with constructing position-over-time graphs appropriate for real-life situations. In the first task, students were asked to use a given diagram to drag blue points and submit three different examples of graphs that represent a girl named Nora’s ride. The ride must meet two requirements: it ends at 13:00 and ends at home. The GeoGebra-based applet (like the left window in figure 2) included a 4-segment interactive graph that can be manipulated by dragging the blue points. The left blue point (a fixed point) represents Nora’s starting location, and the right blue point represents Nora’s finishing location, both relative to home, while the x-axis represents the time relative to 8 am.

In this study, we focused on the second task of the activity (described below), which, unlike the first task, includes a simulation. The uniqueness of the simulation, which adds value to the use of the graph, is that it presents features of the bicycle's motion, such as the change in position relative to home, the change in speed, and the change in direction. These animation features can enhance students' understanding of the underlying mathematics associated with the task. Beyond the given simulation, we reinforced these features through verbal descriptions (including non-critical characteristics in addition to the requirements of the task) that appear in the post-submission report produced automatically according to the data the students chose to submit.

The task

The students were given the following instruction: *Yael rides a bicycle once a week. In each ride, she passes a total distance of 50 km in 4 hours. Use the diagram below, drag the blue points and submit three different examples of graphs representing Yael's ride.*

The task included a GeoGebra-based applet (Figure 2) divided into two windows. The left window includes a 3-segment interactive graph that can be constructed by dragging blue points. The left blue point represents Yael’s starting location, and the right blue point represents Yael’s finishing location, both relative to home, while the x-axis represents the time relative to 8 am. The right window includes a simulation of Yael's ride (overview of the track) corresponding to the graph on the left side of the applet, which is displayed by clicking the play button, and measurements of each simulation state (total distance, starting hour, and total time). Indeed, the applet includes multi-linked representations (MLR), which enable students to investigate various input data by reflecting their actions in a different representation and thus see their own ideas differently.

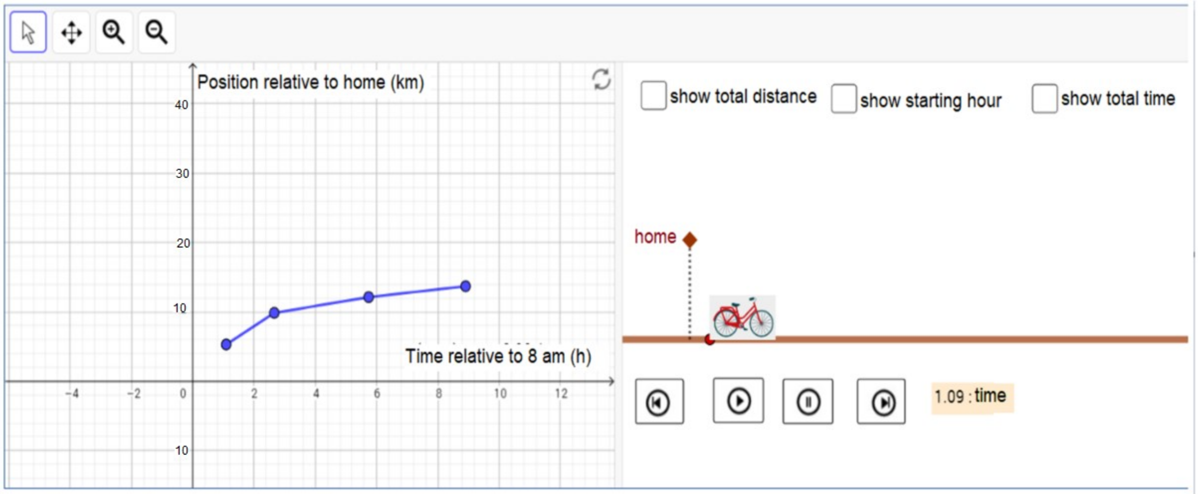


Figure 2: screenshot of the GeoGebra-based applet in task 2

Procedure

The pair of students worked together on the modeling activity within the mathematics class. They solved the tasks in the order in which they appeared in the activity. After completing the first task, the students did not refer to the post-submission report (despite the teacher's request to do so), but they did refer to it after completing the rest of the tasks.

Data collection and analysis

The data collected in the study consisted of (a) students' submissions from the STEP platform, (b) video recordings of the pair's computer screen while working on a modeling activity, and (c) field notes of the researcher taken during the pair observations.

To answer the first research question, the first author attended and observed the whole class, focusing on the pair of students while using the post-submission report. The recordings and notes taken during the observations were then transcribed. We analyzed all the statements and actions that the students performed during the interaction with the post-submission report, while focusing on which parts of the report they referred to. The first author defined and coded the data by mapping each statement and action within the MC and linking it to MMC. She then generated categories for students' statements and actions (detailed below). To answer the second research question, the first author examined the interactions she found while answering the first question and associated them with the modeling process. She formulated the answer to the second question by describing the routes within the MC.

The research was by the Faculty of Education at the University of Haifa and the Israeli Ministry of Education's IRBs.

Findings

Next, we present the submitted examples, the post-submission report of the second task (see Figures 3 and 4), and students' interactions with the post-submission report.

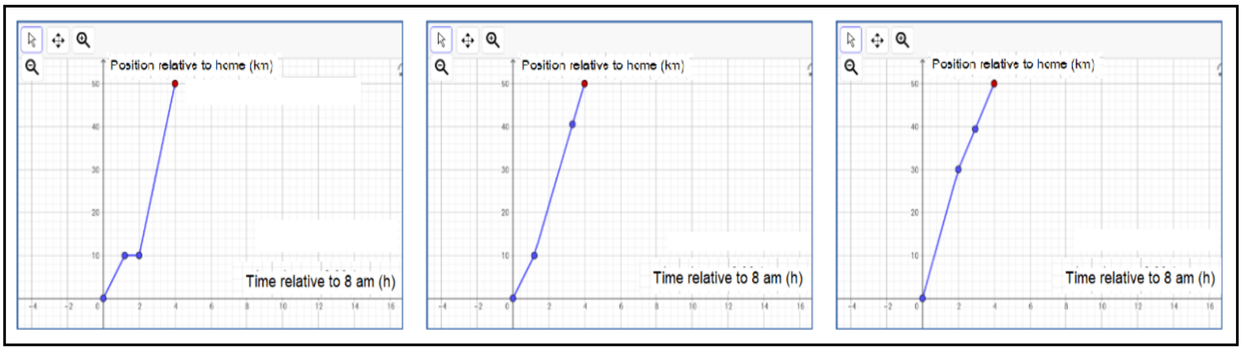
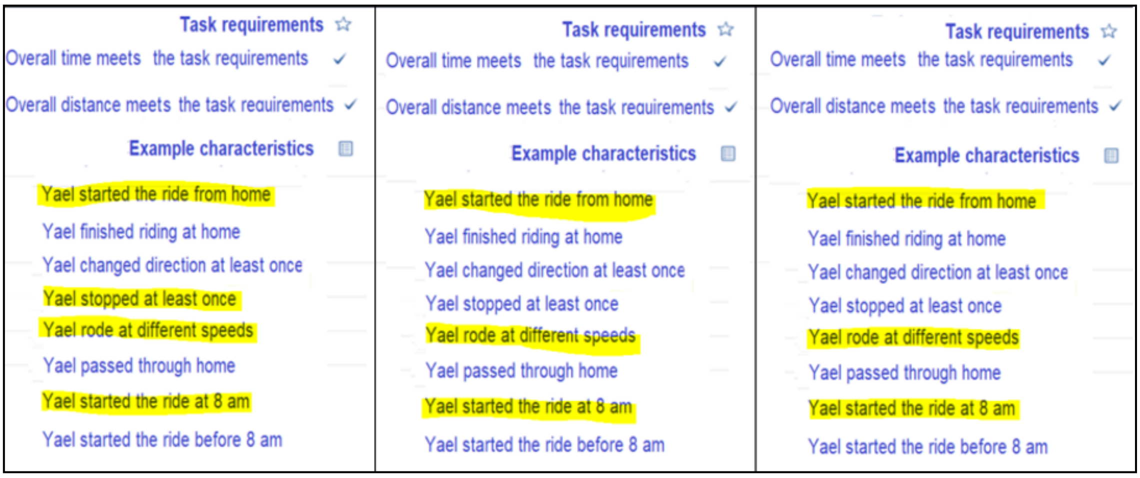

Position relative to home (km)
Time relative to 8 am (h)
Position relative to home (km)
Time relative to 8 am (h)
Position relative to home (km)
Time relative to 8 am (h)

Figure 3: The three examples Ray and Lour submitted

After submitting three examples in the second task (represented in figure 3), Ray and Lour referred to the post-submission report. The post-submission report presented to the students (Figure 4) consisted of two lists that were prepared as part of the task design: (1) Task requirements: a list of critical characteristics that provide information on whether the example is an example of Yael's ride, as described in the task. Each characteristic is marked as V or X according to its existence in the example. (2) Example characteristics: A list of non-critical characteristics that can give students a sense of the breadth of their personal example space. This list includes the characteristics that the task designers would like to deepen the students' interaction with; therefore, the full list appears beneath each example. After analyzing the submitted examples, STEP highlights the characteristics that exist in each example in yellow. Among these characteristics are the various mathematical attributes of the different graphs and their implications on the real-life situation described in the task.

As shown in the report, the submitted examples met the task requirements of overall time and overall distance. In addition, the following eight characteristics were assessed in each example: Yael started the ride from home, Yael finished riding at home, Yael changed direction at least once, Yael stopped at least once, Yael rode at different speeds, Yael passed through home, Yael started the ride at 8 am, and Yael started the ride before 8 am. According to the automated analysis, three characteristics out of them existed in the three examples: Yael started the ride from home, Yael rode at different speeds, and Yael started the ride at 8 am.



Task requirements ☆	Task requirements ☆	Task requirements ☆
Overall time meets the task requirements ✓	Overall time meets the task requirements ✓	Overall time meets the task requirements ✓
Overall distance meets the task requirements ✓	Overall distance meets the task requirements ✓	Overall distance meets the task requirements ✓
Example characteristics ▢	Example characteristics ▢	Example characteristics ▢
Yael started the ride from home	Yael started the ride from home	Yael started the ride from home
Yael finished riding at home	Yael finished riding at home	Yael finished riding at home
Yael changed direction at least once	Yael changed direction at least once	Yael changed direction at least once
Yael stopped at least once	Yael stopped at least once	Yael stopped at least once
Yael rode at different speeds	Yael rode at different speeds	Yael rode at different speeds
Yael passed through home	Yael passed through home	Yael passed through home
Yael started the ride at 8 am	Yael started the ride at 8 am	Yael started the ride at 8 am
Yael started the ride before 8 am	Yael started the ride before 8 am	Yael started the ride before 8 am

Figure 4: the post-submission report Ray and Lour received

Below are some of Ray and Lour's reactions to the post-submission report.

1 Lour: Yael started the ride from home, Yael stopped at least once, Yael rode at different speeds, Yael started the ride at 8 am. [Pointing at the characteristics highlighted in yellow in the list]

2 Ray: Yael finished riding at home. However, they did not tell us. [Pointing at the second characteristic on the list: “Yael finished riding at home,” which appeared beneath all examples].

3 Ray: Yael changed direction at least once. [Pointing at the third characteristic in the list: “Yael changed direction at least once,” which appeared beneath all examples].

4 Lour: Yael stopped at least once. [Pointing at the fourth characteristic in the list: “Yael stopped at least once,” which appears beneath the left example].

5 Lour: Here she stopped. [Pointing at the graph of the left example]

6 Lour: Yael passed through home. Why was this not highlighted? [Pointing at the sixth characteristic in the list: “Yael passed through home,” which appears beneath the three examples].

7 Ray: because we did not make her pass.

8 Ray: all answers are the same. We did a stop only in the left example. [Pointing at the fourth characteristic in the list: “Yael stopped at least once,” which appears beneath the left example].

9 Ray: If she drove for 5 hours then this will be incorrect? [Asking the teacher while pointing at the first task requirement “overall time meet the task requirement”]

10 Teacher: right. You will get X beside the first task requirement

The following categories describe the students’ use of the post-submission report:

Deepening interaction with the example characteristics

The validation process in mathematical modeling involves a series of critical steps aimed at ensuring that the obtained mathematical results accurately reflect real-world situations (Ferri, 2018). In lines 1,4,5 and 8, the students used the verbal description within the report to validate the characteristics of their submitted examples, even though their answers were correct. The interaction with the report enabled them to go through a control process in which the designed characteristics helped them reflect on their answers within the given situation and think about the differences between the graphs submitted. This is not what we would expect in the case of a "correct" answer and indicates the deepening of the interaction with the transition between a realistic situation and a mathematical situation.

The need of external validation

Lines 9-10 describe the interactions between Ray, the report, and the teacher. By asking the teacher about the effect of alternative input data on what is presented in the post-submission report, a “what if” situation, Ray ensures that she understood the first critical characteristic (overall time), and its mathematical representation. This interaction indicates that in some cases, external validation, in this case the teacher’s mediation, is needed in order to understand the results of the report, which is also attributed to the control step within the validation process.

Focusing on the task requirements

In line 2, Ray compared the results of the automated analysis with task requirements. She emphasized that the given situation did not require finishing the ride at home; therefore, they did not submit such an example of a graph. Thus, she validated the mathematical results in a given real situation. This might indicate that students prefer to focus on task requirements more than on other mathematical characteristics.

Unattended characteristics

Ray and Lour did not attend to the non-critical characteristic “Yael started the ride before 8 am” that appears at the end of the list. This might have happened because it was a negation of the characteristic before it, “Yael started the ride at 8 am”.

Next, we describe how students’ use of the post-submission report affects their MMC by describing the new nonlinear modeling route within the MC.

New non-linear modeling route

While associating the analyzed interactions (mentioned above) within the modeling process, we noticed that the interaction with the report, as part of the feedback process, enabled the students to perform a new route within the MC. The students, being at the “real results” phase, arrived to the “real situation” phase by proceeding within the MC in the following order: interpreting the real results presented as verbal descriptions in the post-submission report, validating the mathematical interpretations of the real situation with their submitted examples while using the mathematical model (the graphs), gaining insights to the real situation. For example, in lines 4-5, Lour noticed that the left example represents a situation in which Yael stopped at least once, although the trip description did not require that. Here, she performed a modeling route that included ensuring that the mathematical results obtained accurately reflected the real-world situation by emphasizing what appeared in the mathematical model (by pointing at the graph). Another example is shown in lines 6-7, by pointing at and reading the characteristic "Yael passed through home" Lour attended the real results presented as verbal descriptions. Indeed, this conversation validates the mathematical interpretations of a real situation with the submitted examples while using the mathematical model. In this situation, unlike Lour, Ray recognized that the mentioned characteristic did not exist in their examples, which shows that she made a connection between the mathematical model and the given verbal description. In both examples, the fact that the students made a connection with the given situation can be interpreted as a return to the real situation.

Discussion and conclusions

The aim of the present study was to examine how an automated post-submission report can serve as a bridging tool between the literacy phenomenon and the given mathematical model. To this end, we explored students' use of post-submission reports while working on a modeling activity and its effect on their modeling process. Contrary to traditional mathematical problem solving, in which students often conclude their process with the acquisition of mathematical results, the validation process in mathematical modeling goes beyond (Ferri, 2018). The competencies required for effective validation involve critical checking and reflection on the solutions found. Technological tools can assist in this process not only by providing interactive simulations but also by enhancing the feedback process in other ways. Our findings illustrate this idea and answer the second research question: the verbal descriptions in the post-submission report helped Ray and Lour validate their answers by reviewing various parts of their solution within the given situation and the given mathematical model (graphs) by critically checking the submitted examples and reflecting on them. By interacting with the post-submission report, a non-linear modeling route was identified: from the real result (Phase 6) to the mathematical model (Phase 4) to the real situation (Phase 1). This new route enables students to interpret and validate mathematical results obtained in the extra-mathematical world. These findings support the idea that artifacts such as personal reports, which are designed as tools by others, may become instruments for learning (Yerushalmy et al., 2023) and contribute to the hypothesis that the post-submission report serves as a bridging tool between the literacy phenomenon and the given mathematical model.

Usually, reports are the last step in the feedback process and do not prompt further meaningful interaction with the task, especially if the answer is correct. In this study, the findings, which answer the first research question, show that the post-submission report is an integrative part of the ongoing feedback process, which enables students to make sense of information relating to the task and regarding aspects of their understanding (Hattie & Timperley, 2007). Indeed, looking at the whole interaction of the students with the task as a holistic feedback process that includes using the simulation and interpreting the results of the post-submission report contributes to the validation process that occurs while engaging with the report. For this purpose, a post-submission report is a pivotal part of the task. It should be noted that the findings for this study are preliminary, concerning one pair of students, but we will expand the sample to more pairs.

References

Abu-Raya, K., & Olsher, S. (2021). Learning analytics based formative assessment: Gaining insights through interactive dashboard components in mathematics teaching [Paper presentation]. *AI for Blended-Learning: Empowering Teachers in Real Classrooms Online Workshop*, *EC-TEL'21*, Bozen -Bolzano, Italy.

Cevikbas, M., Kaiser, G., & Schukajlow, S. (2022). A systematic literature review of the current discussion on mathematical modelling competencies: State-of-the-art developments in conceptualizing, measuring, and fostering. *Educational Studies in Mathematics, 109*, 205-236. https://doi.org/10.1007/s10649-021-10104-6

Ferri, R. B. (2018). *Learning how to teach mathematical modeling in school and teacher education.* Springer. https://doi.org/10.1007/978-3-319-68072-9

Greefrath, G. (2011). Using technologies: New possibilities of teaching and learning modelling—Overview. In G. Kaiser, W. Blum, R. Borromeo Ferri, & G. Stillman (Eds.), *Trends in Teaching and Learning of Mathematical Modelling* (ICTMA 14), 301–304. Springer. https://doi.org/10.1007/978-94-007-0910-2_30

Harel, R., Olsher, S., & Yerushalmy, M. (2022). Personal elaborated feedback design in support of students' conjecturing processes. *Research in Mathematics Education, 26(1),* 70-89. https://doi.org/10.1080/14794802.2022.2137571

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research, 77*(1), 81-112. https://doi.org/10.3102/003465430298487

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science, 18*(2), 119-144

Sangwin, C. J., & Köcher, N. (2016). Automation of mathematics examinations. *Computers & Education, 94*, 215-227. https://doi.org/10.1016/j.compedu.2015.11.014

Touma, M., & Olsher, S. (2022). Designing computer-based modeling activities that support students' modeling process. 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),* 1167-1174.

Yerushalmy, M., Olsher, S., Harel, R., Chazan, D. (2023). Supporting inquiry learning: An intellectual mirror that describes what it "sees." *Digital Experiences in Mathematics Education, 9*, 315–342. https://doi.org/10.1007/s40751-022-00120-3

Combining self-assessment and automatic-assessment – a mixed methods study

Carina Tusche[1], Daniel Thurm[1] and Shai Olsher[2]

[1]University of Siegen, Germany; tusche@mathematik.uni-siegen.de , thurm@mathematik.uni-siegen.de

[2]University of Haifa, Israel; olshers@edu.ac.il

This paper introduces a digital learning environment featuring a novel assessment module that merges self-assessment and automatic-assessment. This integration is particularly notable given the limited research on combining these two assessment forms. The module aims to leverage the reflective nature of self-assessment with the efficiency and objectivity of automatic-assessment, potentially enhancing learning outcomes. The paper also outlines the initial phase of a mixed-methods study designed to evaluate this combined assessment module. This study will provide a comprehensive analysis of the module's effectiveness compared to other assessment methods. Through this research, we aim to offer valuable insights into the benefits and limitations of integrating self-assessment and automatic-assessment in educational settings.

Keywords: Automatic-assessment, self-assessment, technology, linear functions.

Theoretical background

Previous studies have shown that digital formative assessment can support student learning, for example through automatic-assessment and feedback on students' solutions (Harel et al., 2022; Olsher & Thurm, 2021). Furthermore, self-assessment is considered important for the development of students' metacognitive skills and the promotion of personal responsibility for their own learning process (Andrade, 2019). However, there is a gap in research that addresses the **combination** of automatic-assessment and self-assessment to support mathematical learning processes (Olsher & Thurm, 2021).

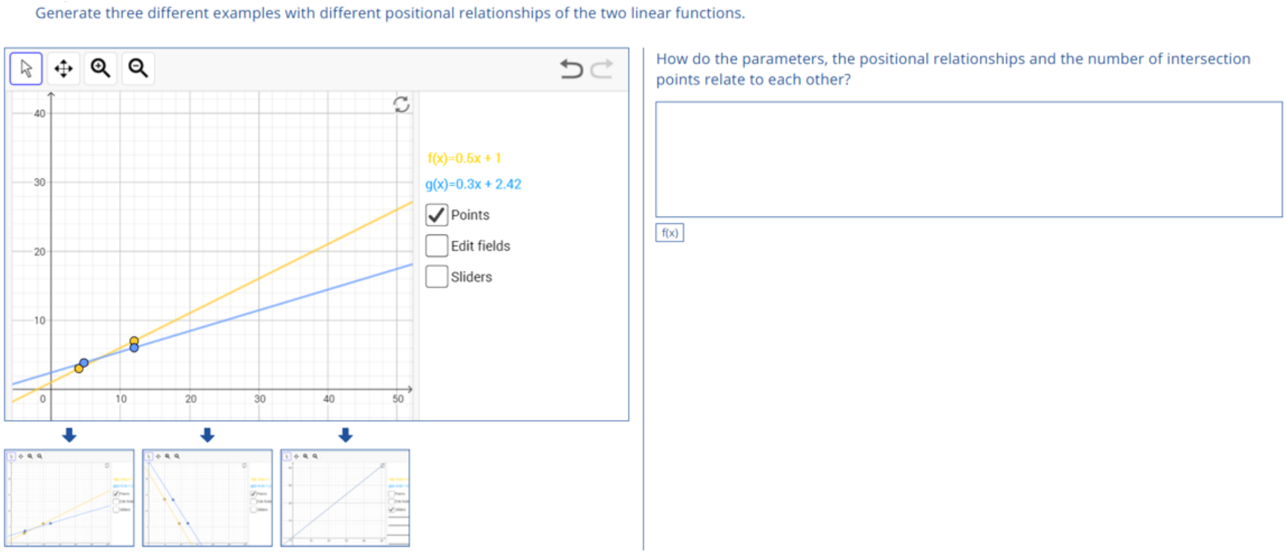
Self-assessment can be conceptualized as a process in which students reflect on the quality of their work, evaluate how well it aligns with established goals or criteria and make modifications accordingly (Andrade et al., 2019). Self-assessment can promote students' metacognitive and self-regulatory processes by encouraging them to evaluate, reflect on and revise their work (Panadero et al., 2017). However, it is important to emphasize that self-assessment carries the risk of students drawing incorrect conclusions about their learning process.

Technology offers various ways to support formative assessment and formative self-assessment, for example with interactive tasks and adaptive real-time feedback (Harel et al., 2022; Olsher & Thurm, 2021, Olsher et al., 2016). As Harel et al. (2022) showed, students working on digital “example-eliciting tasks” (tasks in which students construct examples that illustrate/support their answers to a given problem) can be supported by automatic “attribute isolation elaborated feedback” (AIEF) by providing information on whether specific predefined mathematical characteristics are present in their constructed examples. In this context, Olsher and Thurm (2021) suggested that learner engagement can be further enhanced if they self-assess their work in terms of the predefined characteristics before receiving the AIEF.

Designing an EET addressing the positional relationships of two linear functions

Based on the concept of Olsher and Thurm (2021), we developed a digital learning setting using a GeoGebra applet. The learning setting aims to explore the relationship between the parameters, the number of intersection points, and the positional relationships of two linear functions.

The task requires students to construct three examples with different positional relationships between two linear functions (see Figure 1, left, for one example). To do this, students can move points on the graphs, use sliders or change the parameters on an algebraic level. In the task, students should also formulate a conjecture about how the parameters, the positional relationships and the number of intersection points of the two linear functions relate to each other (Figure 1, box on the right). An example of a students´ conjecture could be: "Both graphs always intersect if one function has a positive and one has a negative slope".


Generate three different examples with different positional relationships of the two linear functions.

How do the parameters, the positional relationships and the number of intersection points relate to each other?

$f(x)=0.5x+1$

$g(x)=0.3x+2.42$

☑ Points

□ Edit fields

□ Sliders

Figure 1: Task and GeoGebra applet

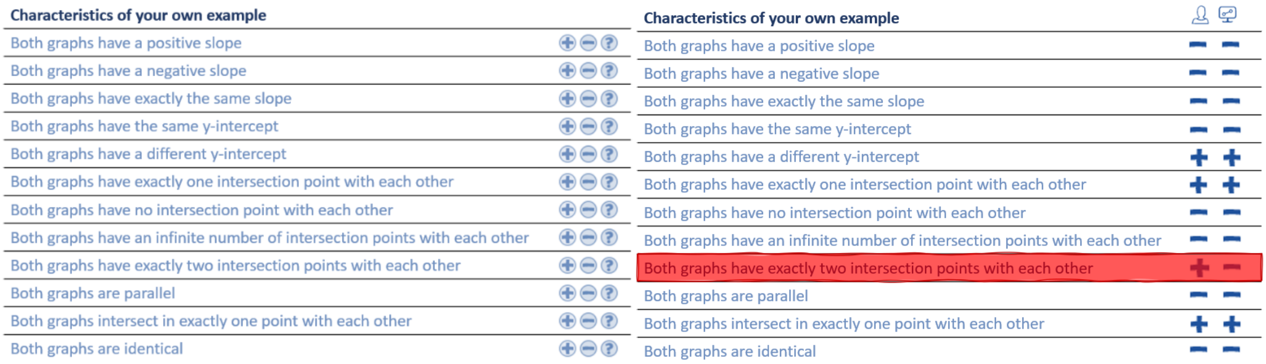
After the students have solved the task (Figure 1), they first evaluate each of their constructed examples and decide which of twelve predefined characteristics are present in their constructed examples (Figure 2, left). If they are unsure whether a characteristic is present, they can indicate this by selecting the question mark. Since the students are supposed to explore the relationship between the parameters, the number of intersection points and the positional relationships, the characteristics were constructed in such a way that each characteristic relates to one of the aspects (parameters, number of intersection points, positional relationships). In addition, we defined a characteristic which is not possible to generate with any example to inspire further reasoning processes. After submitting their work, students receive a report consisting of three parts:

a) an overview of their self-assessment (can no longer be changed),

b) an overview of the automatic-assessment (i.e., overview of which characteristics are present in their examples)

c) a combined overview showing conflicts between the self-assessment and the automatic-assessment (Figure 2, right).

Subsequently students work on the task again to improve their task solution.



Characteristics of your own example	
Both graphs have a positive slope	⊕ ⊖ ?
Both graphs have a negative slope	⊕ ⊖ ?
Both graphs have exactly the same slope	⊕ ⊖ ?
Both graphs have the same y-intercept	⊕ ⊖ ?
Both graphs have a different y-intercept	⊕ ⊖ ?
Both graphs have exactly one intersection point with each other	⊕ ⊖ ?
Both graphs have no intersection point with each other	⊕ ⊖ ?
Both graphs have an infinite number of intersection points with each other	⊕ ⊖ ?
Both graphs have exactly two intersection points with each other	⊕ ⊖ ?
Both graphs are parallel	⊕ ⊖ ?
Both graphs intersect in exactly one point with each other	⊕ ⊖ ?
Both graphs are identical	⊕ ⊖ ?

Characteristics of your own example		
Both graphs have a positive slope	−	−
Both graphs have a negative slope	−	−
Both graphs have exactly the same slope	−	−
Both graphs have the same y-intercept	−	−
Both graphs have a different y-intercept	+	+
Both graphs have exactly one intersection point with each other	+	+
Both graphs have no intersection point with each other	−	−
Both graphs have an infinite number of intersection points with each other	−	−
Both graphs have exactly two intersection points with each other	+	−
Both graphs are parallel	−	−
Both graphs intersect in exactly one point with each other	+	+
Both graphs are identical	−	−

Figure 2: Left: Self-assessment with predefined characteristics; right: the combined overview with highlighted conflicts

Study design

The cluster-randomized mixed-methods study is conducted with approximately 300 ninth grade students divided into three different intervention groups. Within the 45-minute intervention each group engages in the following activities twice:

A) *Combination of self- and automatic-assessment:* This group works on the task, then performs a self-assessment with the characteristics (Figure 1, left) and then receives the combined overview highlighting the conflicts between self-assessment and automatic-assessment (Figure 2, right)

B) *Only automatic-assessment*: This group works on the task, does not carry out a self-assessment, but receives a report with the automatic-assessment (i.e. which characteristics are present in the examples)

C) *Only self-assessment*: This group completes the task and then only carries out the self-assessment with the characteristics, without receiving an automatic-assessment.

To ensure that all intervention groups have the same knowledge at the beginning of the intervention, all three groups receive the same 45-minute introductory phase in which technical terms (e.g. intersection) are learned and repeated. In the introductory phase, no explicit connections are made between the parameters of the linear functions, the number of intersection points and the positional relationships, as these are to be explored in the context of the developed task. At the end of the introductory phase, all students complete a short test to check their knowledge.

Research goal and methods

This study will investigate the extent to which the different interventions affect

- **i)** the metacognitive activities,
- **ii)** the written conjectures,
- **iii)** the variety of generated examples and
- **iv)** the understanding of the relationship between parameters, the number of intersection points and positional relationships of two linear functions.

To reconstruct the metacognitive activities, 4-6 students in each intervention group are filmed, and the recordings are analyzed qualitatively. The written conjectures and the variety of generated examples are reconstructed from the work of the students in the digital learning setting and quantitatively evaluated following the data collection. The understanding of the relationship between parameters, the number of intersection points and positional relationships is determined by a post-test.

Outlook

It is not expected that one of the three interventions will show the best results for all outcome measures i) - iv). By emphasizing the conflicts (between self-assessment and automatic-assessment, see Figure 2, right), intervention A) could help students to use these conflicts for their own learning process. However, the high cognitive load of the combination of self-assessment and automatic-assessment could also be disadvantageous. For example, it may be that the students concentrate more on the reduction of their conflicts and neglect the revision of their hypotheses. In summary, we expect detailed insights into the mathematical learning processes in different assessment conditions, which can help in the design of digital learning settings that integrate automatic-assessment and self-assessment.

References

Andrade H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education*, *4*, 1–13. https://doi.org/10.3389/feduc.2019.00087

Harel, R., Olsher, S., & Yerushalmy, M. (2022). Personal elaborated feedback design in support of students' conjecturing processes. *Research in Mathematics Education*, 1–20. https://doi.org/10.1080/14794802.2022.2137571

Olsher, S., & Thurm, D. (2021). The interplay between digital automatic-assessment and self-assessment. In M. Inprasitha, N. Changsri, & N. Boonsena, (Eds.), *Proceedings of the 44th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, p. 431–440). PME.

Olsher, S., Yerushalmy, M., & Chazan, D. (2016). How might the use of technology in formative assessment support changes in mathematics teaching?. *For the learning of mathematics*, *36*(3), 11–18.

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74–98.

Facilitating teachers' in-the-moment feedback practices

Sumeyra Tutuncu

IOE, UCL's Faculty of Education and Society, London, United Kingdom; s.tutuncu@ucl.ac.uk

This theoretical paper draws attention to the need for further research on the resources that can facilitate mathematics teachers' in-the-moment formative assessment practices, with a focus on the feedback aspect. The paper elaborates on the argument that to effectively integrate formative assessment into their daily teaching practices, teachers need resources that clearly communicate its essence. Thus, I will present an overview of the effects of different types of feedback from literature complemented with a critique of a teacher guide.

Keywords: feedback, formative assessment, curriculum resources.

Introduction

While the key role of formative assessment practices in effective learning has been recognised and acknowledged by policymakers and researchers, it remains a challenge to transfer the essence of formative assessment to teachers' actual in-the-moment practices. This challenge might result from teachers' beliefs regarding the essence of formative assessment as well as a lack of sufficient tools and knowledge (Amado & Morselli, 2023; Antoniou & James, 2014). In that sense, it is important to provide teachers with classroom resources that facilitate a shift in their beliefs as well as provide tools and knowledge.

The five strategies and one big idea of formative assessment proposed by Wiliam and Thompson (2007) can be a useful framework to guide teachers. Namely, they draw attention to the key function of formative assessment as teachers' making evidence-informed decisions and highlight the key strategies as: (1) Clarifying, sharing, and understanding learning intentions and success criteria; (2) Eliciting evidence of learning; (3) Providing feedback that moves learning forward; (4) Activating learners as instructional resources for another; (5) Activating learners as owners of their learning.

In this paper, the third strategy is the key focus. After presenting an overview of the various types of feedback and their potential influence on students' learning under various conditions, in order to complement the theoretical discussions, I will critique a teacher guide that provides explicit recommendations for feedback. This paper contributes new insights into the potential role of daily curriculum resources in challenging teachers' ineffective feedback practices.

Varying impact of different types of feedback on learning in different situations

The term feedback was originally used to refer to the information that is provided to the learner to reduce the gap between the actual learning and intended learning (Ramaprasad, 1983). More recent conceptualisations considered feedback as a process in which students are expected to be engaged with the information to improve their work (e.g., Hattie & Timperley, 2007).

Intriguingly, the research evidence points to unfavourable potential consequences of teachers' feedback on students' learning (Kluger & DeNisi, 1996). Among the studies Kluger and DeNisi reviewed in their comprehensive review, more than one-third of the feedback interventions harmed students' learning mainly due to the attention on student self instead of the task-related feedback.

More recent reviews provided further insights into the conditions that result in positive or negative effects of feedback (Hattie & Timperley, 2007; Shute, 2008; Van der Kleij et al., 2015).

These three reviews provided insights into the impact of different types of feedback on learning. Hattie and Timperley suggested four levels of feedback: whether the task was accomplished (task-level); how to accomplish task or how to improve the intended product (process-level); highlighting students' skills as learner (self-regulation level) and personal comments on students' personality (self-level). Similar to Kluger and DeNisi's findings, Hattie and Timperley's review has taken attention to the harmful effect of feedback that focuses on the self. Further, feedback types can be categorised according to the extent to which the teacher provides information about students' response: telling whether the answer is correct or incorrect; providing the correct answer; and providing additional instructions with elaborated feedback (Shute, 2008). Another common categorisation of feedback types is on timing: feedback can be either provided immediately or it can be delayed (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Shute, 2008).

Such variety within the feedback practices can suggest a challenge for teachers to decide which type can be helpful in which situation in the classroom. Existing research provides insights into the impact of various feedback types according to students' prior learning or the level of skill the task requires. More explicitly, while immediate feedback can be useful for enabling students' engagement with difficult tasks in the early steps and consolidating procedural or conceptual knowledge, delaying feedback can prevent students' potential distraction when they are working on easy tasks and can be beneficial for long-term learning of high-level skills (Fyfe & Rittle-Johnson, 2016; Shute, 2008).

It should be noted that most of the studies that provide insights into the effect of feedback were conducted in an artificial experimental design environment beyond reflecting on teachers' real classroom feedback practices. Antoniou and James (2014) explored primary school teachers' existing tendencies related to formative assessment practices. The findings of their study suggest that teachers tend to provide immediate corrective feedback and judgment rather than providing feedback that can enhance students' learning. Additionally, they found short-term rewards and the use of grades as common feedback strategies, which may undermine students' intrinsic motivation and hinder their long-term development.

To conclude, the existing literature provides insights into the various impacts of different types of feedback on students' learning. Also, there is evidence that teachers may tend to not pay attention to these different types of feedback but they may provide rewards that can be an example of the feedback for self and immediate corrective feedback, which may not ultimately support long-term and high-level learning. One way of facilitating teachers' change can be the development of teachers' daily resources in which effective feedback practices are integrated.

A critique of a teacher guide that involves explicit suggestions for feedback

In the previous section, I provided an overview of the types of feedback and their potential impact on learning as well as highlighting teachers' need for daily use resources that can facilitate their feedback practices. In this section, employing the insights gained from this literature, and drawing on the analytical strategies developed earlier (Tutuncu et al., 2023), I will critique a teacher guide for its integrating of specific guidance for teachers' practices. This teacher guide is from a set of materials on the Mathematics Assessment Project (MAP). This teacher guide was chosen as it provides rich insights and tools for teachers for formative assessment practices in general and also involves specific guidance for feedback practices. However, despite its strong focus on providing guidance and tools for teachers, the provision of guidance and tools for feedback is limited to only two aspects. In the following paragraphs, these examples are elaborated.

First, in the teacher guide, teachers were provided with guidance for process-level and immediate feedback when students struggled to start working on an open answer and open-method task, which could enable students to understand what was required in the task. Teachers were provided with detailed guidance in terms of the way of providing this process-level and immediate feedback. Importantly, they were advised to use questioning rather than directly telling them what to do, such as the following example questions.

> What useful information are you given? Underline this. What do you need to find out? How can you use the information you know to do this? (MAP, Drawing to Scale: A garden, page T-4)

This immediate feedback that was provided in the initial steps of students' work could facilitate students' engagement with the rest of the task as Shute (2008) suggested. In the literature, one limitation of immediate feedback is suggested as hindering high-level learning. The suggested immediate feedback in this teacher guide could mitigate this risk. That is to say, using prompts that assist students to realise what procedures they should apply and connect this understanding with their existing knowledge might support students' strategic competence and adaptive reasoning beyond procedural fluency and conceptual understanding (Kilpatrick et al., 2001).

Second, teachers were recommended against providing evaluative feedback. The rationale for this advice was explained with students' potential distraction by comparing their scores with their peers' scores rather than focusing on how to improve their learning. As an alternative to scoring, teachers were encouraged to use questioning to enable students to reflect on their work. As an example, when teachers observed that students have difficulty in making calculations with decimals, they were advised to use the following questions as feedback.

> How do you convert 1m in real life to a measurement on the plan? What about 3m? Now apply the same method to figure out what the length 3.25 m is on the plan. (MAP, Drawing to Scale: A garden, page T-4)

These questions can be more beneficial than traditional immediate-corrective or immediate-elaborated feedback. That is to say, rather than correcting students or re-teaching the topic, students are encouraged to reflect on their work by using numbers with which the calculations can be relatively straightforward.

Conclusion

This paper draws attention to the need for teachers' daily resources that have the potential to facilitate teachers' effective in-the-moment feedback practices. The existing literature offers insights into the impacts of different types of feedback on students' learning. Although studies exclusively examining teachers' daily feedback practices without any form of intervention are constrained in their scope and potential insights, there is evidence that teachers tend to provide immediate feedback which is corrective or judgmental rather than constructive forms of feedback. In this paper, I present an example teacher guide that has the potential to facilitate mathematics teachers' more effective feedback practices, drawing attention to the limited existing resources that can have this potential.

References

Amado, N., & Morselli, F. (2023). Teachers' beliefs about assessment: A study in Italy and Portugal. In Drijvers, P., Csapodi, C., Palmér, H., Gosztonyi, K., & Kónya, E. (Eds.). *Proceedings of the Thirteenth Congress of the European Society for Research in Mathematics Education (CERME13)* (pp 3916–3923). Alfréd Rényi Institute of Mathematics and ERME, Budapest, Hungary.

Antoniou, P., & James, M. (2014). Exploring formative assessment in primary school classrooms: Developing a framework of actions and strategies. *Educational Assessment, Evaluation and Accountability*, *26*(2), 153–176. https://doi.org/10.1007/s11092-013-9188-4

Fyfe, E. R., & Rittle-Johnson, B. (2016). Feedback both helps and hinders learning: The causal role of prior knowledge. *Journal of Educational Psychology*, *108*(1), 82–97. https://doi.org/10.1037/edu0000053

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding It Up: Helping Children Learn Mathematics*. Washington, DC, USA: National Academies Press.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, *28*(1), 4–13. https://doi.org/10.1002/bs.3830280103

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Tutuncu, S., Hodgen, J. & Golding, J. (2023). Proposing educative features of curriculum materials that can enhance teachers' noticing. In Drijvers, P., Csapodi, C., Palmér, H., Gosztonyi, K., & Kónya, E. (Eds.). *Proceedings of the Thirteenth Congress of the European Society for Research in Mathematics Education (CERME13)* (pp. 4205–4212). Alfréd Rényi Institute of Mathematics and ERME, Budapest, Hungary

Van der Kleij, F., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A Meta-analysis. *Review of Educational Research*, *85*(4), 475–511. https://doi.org/10.3102/0034654314564881

Wiliam, D., & Thompson, M. (2007). Integrating assessment with learning: What will it take to make it work? In C. A. Dwyer (Ed.), *The Future of Assessment* (pp. 53–82). Mahwah, NJ, USA: Lawrence Erlbaum Associates.

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.

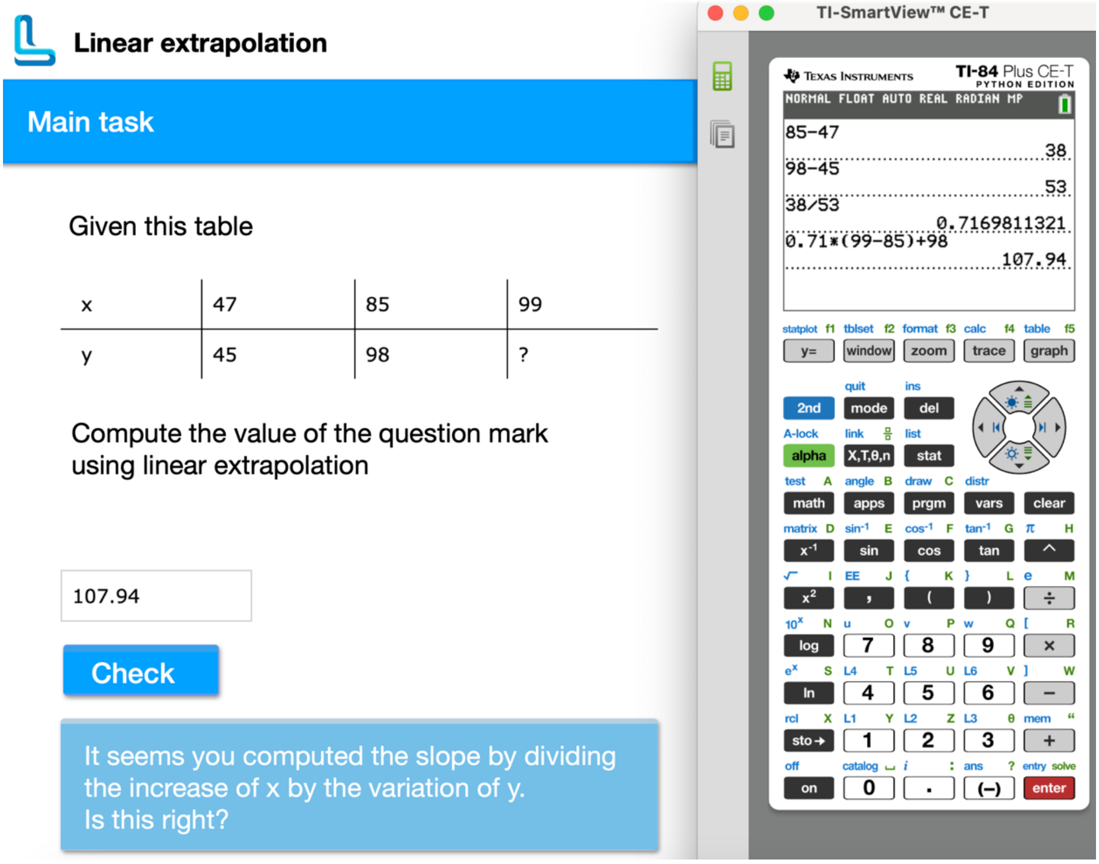
Subtask C: Calculate slope.

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.



Linear extrapolation

Main task

Given this table

x	47	85	99
y	45	98	?

Compute the value of the question mark using linear extrapolation

107.94

Check

It seems you computed the slope by dividing the increase of x by the variation of y.
Is this right?

```
NORMAL FLOAT AUTO REAL RADIAN MP
85−47
                                  38
98−45
                                  53
38/53
                        0.7169811321
0.71*(99−85)+98
                              107.94
```

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].

GeoGebra pop-up notifications as formative feedback for learning computational thinking in mathematics lessons

Wahid Yunianto[1], Theodosia Prodromou[2] and Zsolt Lavicza[3]

[1]Johannes Kepler University Linz, School of Education, Linz, Austria; yunianto.wah@gmail.com

[2]University of New England, Australia; theodosia.prodromou@une.edu.au

[3]Johannes Kepler University Linz, School of Education, Linz, Austria; zsolt.lavicza@jku.at

Computational thinking (CT) is a fundamental skill for everyone that is relevant to 21st-century skills. Initiatives have been carried out to enhance CT outside of computer science (CS) courses. In this study, we integrated CT into mathematics lessons utilizing GeoGebra. Our lessons' development was guided by the educational design research approach (EDR). When we piloted our lessons with junior high school students, we found that GeoGebra could provide pop-up notifications as formative feedback to support students' debugging skills. Pop-up notifications could act as a negotiated-style interruption or an immediate-style interruption. The first interruption is when the students get a pending message about the errors so that they become aware of them. Meanwhile, the latter informs students to revise the errors immediately. In this paper, we will discuss how these interruptions as a formative assessment could be a means of support for students while learning CT+Maths lessons.

Keywords: Computational thinking, mathematics, formative feedback, assessment, GeoGebra.

Introduction

The purpose of this study is to explore formative feedback that can be supported by GeoGebra while learning computational thinking in mathematics lessons. Papert (1980) introduced computational thinking (CT) through computer programming within mathematics contents. His idea was to build students' CT skills and mathematical knowledge when interacting with the LOGO turtle. A program was developed by him and his colleagues. Students could input the commands and observe what happened to them. For instance, to make a triangle, students could type in: forward 50, right 120, forward 50, right 120, forward 50, and right 120. It would produce an isosceles triangle with a length of 50 for each vertice. The program was intended to be intuitive, and students could develop CT skills as well as mathematics concepts, which were considered too formal for young learners.

Since LOGO Turtle and its similar programs were introduced, schools in the US have started to initiate programming activities. However, Resnick (2009) found that this initiative was not sustainable due to some challenges, such as students finding it difficult with the textual commands (syntaxes). Later, he and his colleagues developed what we now call block programming, 'Scratch' to smoothen how students learn programming. Nowadays, Scratch is used by users around the world. Additionally, in order to enhance CT in mathematics learning, visual programming (Scratch) and mathematics software (spreadsheets, GeoGebra, and MATLAB) have been utilized (Ye et al., 2023). However, visual programming tends to provide little space for debugging as it was not designed to do so (Liu et al., 2017), and block programming prevents syntax errors from happening (Resnick, 2012). Therefore, our lessons utilized GeoGebra, which has a 'pop-up' feature to facilitate debugging when command errors are inputted.

Ukkonen (2023) asserted that integrating CT in mathematics education is new, and to ensure successful integration, we need to consider formative assessment. Instead of the formative feedback provided by the teachers, we consider formative feedback from the tool students use (GeoGebra) that aligns with the idea of constructionism by Papert (1980) in which students build up their knowledge through their interaction with computer programs. We would like to present our experience when utilizing GeoGebra for formative feedback while learning CT in mathematics lessons. Therefore, we would like to find out how GeoGebra's pop-up could support students in accomplishing debugging tasks, which is a form of formative assessment.

Conceptual Framework

We commonly have two types of assessment, namely formative (assessment for learning) and summative (assessment of learning) (Tan, 2011). The first aims to support students to reach their learning goals through feedback during the learning process and enhance students' learning (Black & Wiliam, 1998). For instance, when a student had a misconception about adding a decimal number with an integer by neglecting the decimal separator, the teacher could assist students with questions and/or feedback to deal with the misconception. Meanwhile, the second is conducted at the end of the topic, lesson, term, or semester to know what students have gained from their learning. In this paper, we are interested in formative assessment, which mainly focuses on formative feedback.

Formative feedback

Cizek et al. (2019) defined formative feedback as teachers' and students' support to infer strengths, weaknesses, and opportunities for improvements in learning. They stated that with formative assessment, students could deepen their understanding, improve their achievement, take responsibility for their learning, and self-regulate their learning. However, Burkhardt and Schoenfeld (2019) argued that it is unlikely that teachers could provide formative assessments for each student during a class teaching period because of time limitations.

As formative assessment could be in the form of feedback, Hattie and Timperley (2007) explained how the power of feedback could benefit learning. Feedback is information that leads students to confirm, add to, overwrite, tune, or restructure it in memory (Winnie & Butler, 1994 as cited in Hattie & Timperley, 2007). Hattie and Timperley (2007) described three directions and four levels of feedback. For the directions, they classified them into (1) feedback about where a student is (FeedBack), (2) where a student is going (FeedUp) and (3) where a student will go next (FeedForward). The following model (Figure 1) by (Hattie & Timperley, 2007, p. 87) is used to help us consider the feedback provided by GeoGebra pop-up.

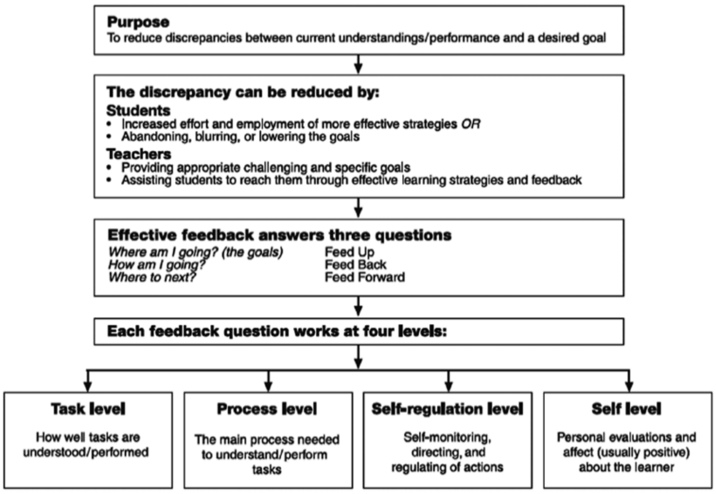

Purpose
To reduce discrepancies between current understandings/performance and a desired goal
The discrepancy can be reduced by:
Students
• Increased effort and employment of more effective strategies OR
• Abandoning, blurring, or lowering the goals
Teachers
• Providing appropriate challenging and specific goals
• Assisting students to reach them through effective learning strategies and feedback
Effective feedback answers three questions
Where am I going? (the goals)
How am I going?
Where to next?
Feed Up
Feed Back
Feed Forward
Each feedback question works at four levels:
Task level
How well tasks are understood/performed
Process level
The main process needed to understand/perform tasks
Self-regulation level
Self-monitoring, directing, and regulating of actions
Self level
Personal evaluations and affect (usually positive) about the learner

Figure 1: A model of feedback to enhance learning (Hattie & Timperley, 2007, p. 87)

Ukkonen (2023) used this model for analyzing a case study with a teacher and two students when learning computational thinking in a mathematics lesson. In this paper, we will utilise this model to specify to which parts the GeoGebra pop-ups belong to the model, discuss, and analyse how the GeoGebra pop-ups have helped students in our study.

Technology and Formative Assessment

Digital technology can gather and process data on a large scale and speed, making it a powerful resource to support teachers in formative assessment (Looney, 2010). Dalby and Swan (2019) argued that digital technology could support and empower students in formative processes by providing feedback. Additionally, the feedback could act as a replacement for or an addition to teacher-led processes (Dalby & Swan, 2019). From their study, we could learn that technology could help process students' responses quickly and provide feedback quickly so that students can modify their responses.

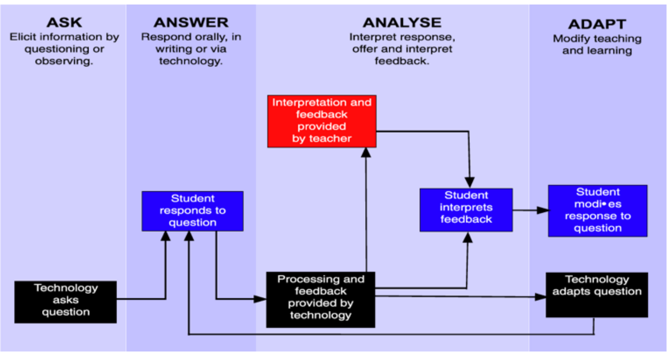

ASK
Elicit information by questioning or observing.
ANSWER
Respond orally, in writing or via technology.
ANALYSE
Interpret response, offer and interpret feedback.
ADAPT
Modify teaching and learning
Interpretation and feedback provided by teacher
Student responds to question
Student interprets feedback
Student modi• es response to question
Technology asks question
Processing and feedback provided by technology
Technology adapts question

Figure 2: A model of feedback to enhance learning (Dalby & Swan, 2019, p. 841).

The AuthoMath project (https://www.authomath.org/) utilizing STACK and GeoGebra is relevant to our study. Using STACK and GeoGebra for formative assessment has helped students to be more successful with accomplishing the tasks (Sanz-Ruiz et al., 2023). With this digital technology, students would receive feedback if they made mistakes by pressing a check button. The difference from our study is that GeoGebra will automatically notify students when incorrect commands are inputted without pressing the check button. Our previous study, Yunianto et al. (2023) utilized this GeoGebra feature (pop-up notifications), and it benefited students. GeoGebra's pop-up is related to two styles of interruptions, namely, the immediate style of interruption and the negotiated style of interruption. Robertson et al. (2004) investigated these interruptions when students were debugging and found that negotiated-style interruption is better than immediate-style interruption when supporting debugging activities.

Method

This study explored the potential of the GeoGebra pop-up as formative feedback when learning computational thinking (CT) in mathematics lessons. It investigated how GeoGebra pop-up could act as formative feedback to benefit students when accomplishing Math+CT tasks. The analysis of our data about how this feature has helped us in our study is qualitative descriptive research as it provides a comprehensive summary of data describing the primary properties of the participants' actions, words, and experiences during data collection and then interpreting it (Ayton, 2023).

Context

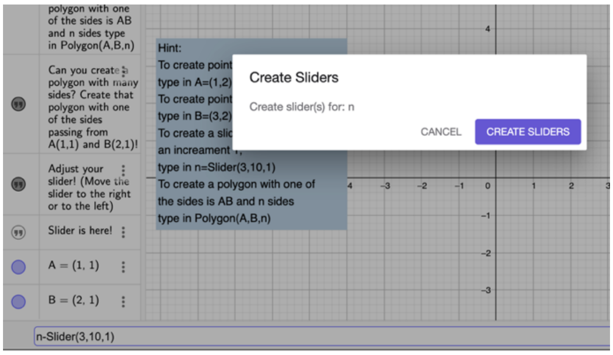
We developed GeoGebra-based Math+CT lessons on the topic of the area of a circle (Yunianto et al., 2023) guided by an educational design research (EDR) approach (McKenney & Reeves, 2018). More details of the development of GeoGebra-based Math+CT lessons can be found in our previous study. In the lessons, students learned how to construct an inscribed regular polygon on a circle with GeoGebra commands. Students also had to debug a program to create a circle and an inscribed polygon. The participants of this study were seventeen junior high school students (aged 12-15) from five different schools in Indonesia.

Data collection and analysis

We collected data by recording students' screens while they were engaging in the tasks. Students were well informed by their teachers about this research and consented to participate in this study. Not all videos recorded full lessons due to the limited storage capacity of the devices and technical issues. Some videos captured the voices of students and a teacher. For this paper, we could not provide the analysis of all videos but rather a few videos that depict the potential of formative feedback. We transcribed students' discussions with themselves, their peers, and the teachers and then provided the interpretation of the transcripts. The analysis was one of progressive focussing (Robson, 1993). At the first stage, the recordings were simply transcribed, and screenshots were incorporated as necessary to make sense of the transcription. Subsequently, the first author turned the transcript into a plain account and an interpretative account was written and the validity of those interpretation was discussed with other authors.

Findings

The following paragraphs present the work of a student who worked on constructing a regular polygon with a slider. The polygon vertices could change as the slider was moved (from 3 to 10). Therefore, this construction would produce an equilateral triangle, a square, a pentagon, a hexagon, a heptagon, an octagon, a nonagon, and a decagon respectively to the slider. This student inputted an incorrect command for the slider 'n-Slider(3,10,1)' and GeoGebra notified her with a pop-up. It said, "Create slider(s) for n". She then edited the command to be n=Slider(3,10,1) and continued to input the next command. Finally, she accomplished this task by creating a regular polygon with a slider.



polygon with one of the sides is AB and n sides type in Polygon(A,B,n)

Can you create a polygon with many sides? Create that polygon with one of the sides passing from A(1,1) and B(2,1)!

Adjust your slider! (Move the slider to the right or to the left)

Slider is here!

A = (1, 1)

B = (2, 1)

Hint:
To create point
type in A=(1,2)
To create point
type in B=(3,2)
To create a slic
an increament
type in n=Slider(3,10,1)
To create a polygon with one of
the sides is AB and n sides
type in Polygon(A,B,n)

Create Sliders

Create slider(s) for: n

CANCEL CREATE SLIDERS

n-Slider(3,10,1)

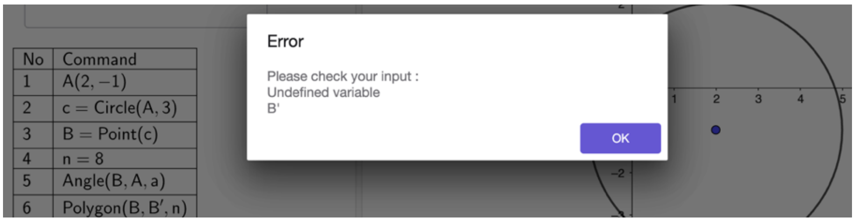
Figure 3: A pop-up notification for an incorrect slider command

While doing this task, this student talked to herself immediately after the pop-up notification appeared. The self-talk of this student is as follows:

Student A: Why is it incorrect? (the student was thinking for a while)
Student A: (self reply) Oh this one, the equal sign.

From the transcript, we can see that this student reflected on her command after seeing the pop-up notification and thought for a while what was wrong. Her self-reflection led her to the correct answer.

On another task, she also benefitted from the notification. She was working on a task and inputted all correct commands but missing a (the angle). The pop-up (Figure 4) notified the error (*Please check your input, undefined variable B*'). The student misinterpreted the notification first and she edited the polygon, thinking it was the source of error. Afterwards, she discussed with her peer, and they reflected to the previous experiences of similar tasks. Based on their prior knowledge and peer discussion, she realized that she missed the a. Hence, she inserted a (a=360deg/n). In the end, she had 3 runs and 0 clear and accomplished the task.



No	Command
1	A(2, −1)
2	c = Circle(A, 3)
3	B = Point(c)
4	n = 8
5	Angle(B, A, a)
6	Polygon(B, B′, n)

Error

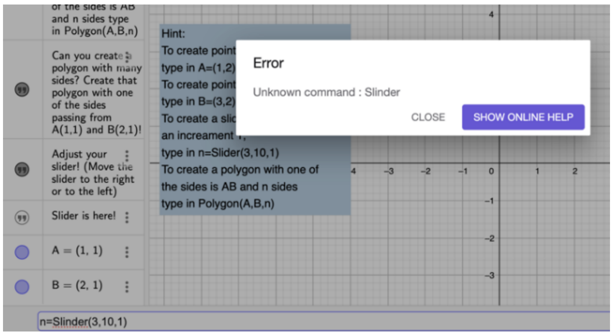
Please check your input :
Undefined variable
B'

OK

Figure 4: A pop-up notifying an error

The following excerpts present another case when a student discussed with his teacher when he encountered the pop-ups. He inputted commands for point A and point B correctly. He typed in n=Slinder(3,10,1), with the additional letter n. A pop-up appeared notifying the "Unknown command: Slinder" (Figure 5). He edited it and pressed enter, then he deleted it.

Student B: Why is it incorrect?
Teacher: That is there to make a slider.
Student B: I saw it from the hint!
Teacher: Yes, you can see it from the hint
Student B: I made it sir!
Teacher: Are you sure? Where is the slider?
Student B: Look at this first!(pointing to the slider on screen)



of the sides is AB
and n sides type
in Polygon(A,B,n)

Can you create a
polygon with many
sides? Create that
polygon with one
of the sides
passing from
A(1,1) and B(2,1)!

Adjust your
slider! (Move the
slider to the right
or to the left)

Slider is here!

A = (1, 1)

B = (2, 1)

Hint:
To create point
type in A=(1,2)
To create point
type in B=(3,2)
To create a slic
an increament
type in n=Slider(3,10,1)
To create a polygon with one of
the sides is AB and n sides
type in Polygon(A,B,n)

Error

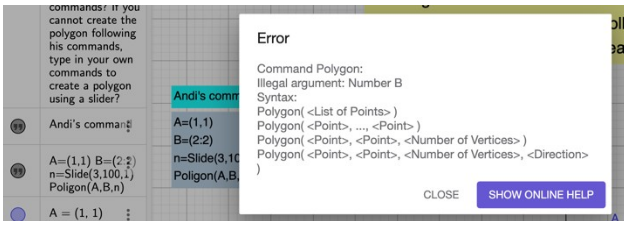
Unknown command : Slinder

n=Slinder(3,10,1)

Figure 5: A pop-up notification as an immediate-style interruption

The above excerpts show how the GeoGebra pop-up notification provided immediate feedback to the student, so he looked at the hint to receive guidance about how to create a slider.

In another case, we witnessed a negotiated style of interruption where an incorrect point B command B=(2:2) was inputted but did not provide a pop-up because it could be run as a slider. This notification was delayed until it was used for the polygon's command. It would notify it as incorrect when it is used for the polygon command as B was not a point (*illegal argument Number B*) (Figure 6). It helped students to locate the error so that they could go directly and fix the command that caused the error. Besides it directed students to the error, it suggested them how to use correct commands and variables when programming geometrical objects. For instance, in this case they only needed to use points to create a polygon.



commands? If you
cannot create the
polygon following
his commands,
type in your own
commands to
create a polygon
using a slider?

Andi's command

A=(1,1) B=(2:2)
n=Slide(3,100,1)
Poligon(A,B,n)

A = (1, 1)

Andi's comm

A=(1,1)
B=(2:2)
n=Slide(3,10
Poligon(A,B,

Error

Command Polygon:
Illegal argument: Number B
Syntax:
Polygon(<List of Points>)
Polygon(<Point>, ..., <Point>)
Polygon(<Point>, <Point>, <Number of Vertices>)
Polygon(<Point>, <Point>, <Number of Vertices>, <Direction>
)

Figure 6: A pop-up notification as a negotiated-style interruption

We counted the number of pop-up cases on our 17 videos. In total, there are 35 cases where 27 of them led students to successfully accomplish the tasks because both styles of interruption informed them about the errors and how they could fix them. Nevertheless, we did not investigate if the students struggle with the pop-ups that were written in English language.

Discussion

Students benefited from the GeoGebra pop-up notifications that helped them to complete the tasks successfully. This result is in line with the study by Sanz-Ruiz et al. (2023). We observed that in some cases after our students received a pop-up, they made fewer errors when they engaged later with the remaining tasks. This might be called effective error detection skills as per Hattie and Timperley (2007). Students experienced errors and would notice similar pop-ups if they made the same mistakes. They could avoid the same mistakes in later tasks. In this paper, we studied the impact of GeoGebra pop-up notifications on students' actions at the task level of the model of Figure 1. In future studies, we aspire to examine in detail the impact of feedback on students' actions and interactions at the process level, self-regulation level, and self-level.

We showed that the formative feedback from the GeoGebra pop-ups provided students with immediate feedback and notification about their errors so they could reflect on their actions, revise, and make relevant corrections. This type of feedback belongs to the Feed Back direction at first level, feedback of the task (FT) and second level, feedback of the process (FP) by Hattie and Timperley (2007). The provided information about pop-ups and the task is similar to the study by Simmons and Cope (1993) in which LOGO feedback was attributed to FT leading students to more correct answers. At the FP level of the model (Figure 1), the GeoGebra pop-up notification (Figure 6) could act as a cueing mechanism (Hattie & Timperley, 2007),which helps students improve their strategies of programming and coding. It could be in the form of immediate-style of interruption and negotiated-style of interruption to support students reflect on their mistakes. Our data shows that students spent more time to identify their errors when they used the negotiated style of interruption. On the contrary, the immediate-style interruption informed them directly about their errors. However, our data do not support that immediate style of interruption is better than negotiated style of interruption as suggested by the literature (Robertson et al., 2004). Our data show that both styles of interruptions are beneficial to students.

Our GeoGebra-based Math+CT lessons seem to follow the model in Figure 2 by Dalby and Swan (2019). We provided students with GeoGebra tasks. If students input incorrect commands, a pop-up will notify them. Students could proceed with their actions or discuss the pop-up with their peers and teacher. Next, students could delete or edit their incorrect commands and proceed with the next task.

References

Ayton, D. (2023). Qualitative Descriptive Research. In D. Ayton, T. Tsindos, & D. Berkovic (Eds.), *Qualitative Research – a practical guide for health and social care researchers and practitioners* (pp. 49–54). Monash University.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *International Journal of Phytoremediation*, *21*(1). https://doi.org/10.1080/0969595980050102

Burkhardt, H., & Schoenfeld, A. (2019). Formative Assessment in Mathematics. In H. L. Andrade, R. E. Bennett, & G. J. Cizek (Eds.), *Handbook of Formative Assessment in the Disciplines* (pp. 35–67). Taylor and Francis. https://doi.org/10.4324/9781315166933-3

Cizek, G. J., Andrade, H. L., & Bennett, R. E. (2019). Formative Assessment: History, Definition, and Progress. In *Handbook of Formative Assessment in the Disciplines*.

Dalby, D., & Swan, M. (2019). Using digital technology to enhance formative assessment in mathematics classrooms. *British Journal of Educational Technology*, *50*(2). https://doi.org/10.1111/bjet.12606

Hattie, J., & Timperley, H. (2007). The power of feedback. In *Review of Educational Research* (Vol. 77, Issue 1). https://doi.org/10.3102/003465430298487

Liu, Z., Zhi, R., Hicks, A., & Barnes, T. (2017). Understanding problem solving behavior of 6–8 graders in a debugging game. *Computer Science Education*, *27*(1). https://doi.org/10.1080/08993408.2017.1308651

Looney, J. (2010). Making it Happen: Formative Assessment and Educational Technologies. *Promethean Thinking Deeper Research Papers*.

McKenney, S., & Reeves, T. C. (2018). Conducting Educational Design Research. In *Conducting Educational Design Research*. https://doi.org/10.4324/9781315105642

Papert, S. (1980). *Mindstorms: Children, Computers, and Powerful Ideas* (1st ed.). Basic Books.

Resnick, M. (2012). Reviving Papert's Dream. *Educational Technology*, *52*(4).

Resnick, M., Maloney, J., Monroy-Hernández, A., Rusk, N., Eastmond, E., Brennan, K., Millner, A., Rosenbaum, E., Silver, J., Silverman, B., & Kafai, Y. (2009). Scratch: Programming for all. *Communications of the ACM*, *52*(11). https://doi.org/10.1145/1592761.1592779

Robertson, T. J., Prabhakararao, S., Burnett, M., Cook, C., Ruthruff, J. R., Beckwith, L., & Phalgune, A. (2004). Impact of interruption style on end-user debugging. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/985692.985729

Robson, C. (1993). *Real World Research: A Resource for Social Scientists and Practitioner-Researchers* (1st ed.). Blackwell.

Sanz-Ruiz, M., Diego-Mantecón, Manuel, J., & Ortiz-Laso, Z. (2023). Formative feedback on linear equations. *CADGME 2023: Digital Tools in Mathematics Education*.

Simmons, M., & Cope, P. (1993). Angle and rotation: Effects of different types of feedback on the quality of response. *Educational Studies in Mathematics*, *24*(2). https://doi.org/10.1007/BF01273690

Tan, K. (2011). Assessment for Learning Reform in Singapore – Quality, Sustainable or Threshold? In *Assessment Reform in Education*. https://doi.org/10.1007/978-94-007-0729-0_6

Ukkonen, A. (2023). Formative assessment of computational thinking in mathematics – a case study. *Proceedings of the Thirtieth Congress of the European Society for Research in Mathematics Education*, 1–8.

Ye, H., Liang, B., Ng, O.-L., & Chai, C. S. (2023). Integration of computational thinking in K-12 mathematics education: a systematic review on CT-based mathematics instruction and student learning. *International Journal of STEM Education*, *10*(3), 1–26. https://doi.org/10.1186/s40594-023-00396-w

Yunianto, W., Prodromou, T., & Lavicza, Z. (2023). Debugging on GeoGebra-based Mathematics+Computational Thinking lessons. *28th Asian Technology Conference in Mathematics*, 322–331.

Mathematics teacher educators' microteaching implementations: selection of settings and approaches to assessment and feedback for prospective primary teachers

Bilge Yılmaz Aslan[1] Mehmet Fatih Özmantar[2] Bilge Kuşdemir Kayıran[3]

[1]Gaziantep University, Faculty of Education, Gaziantep, Türkiye; bilqe.yilmaz@gmail.com

[2]Gaziantep University, Faculty of Education, Gaziantep, Türkiye; ozmantar@gantep.edu.tr

[3]Gaziantep University, Faculty of Education, Gaziantep, Türkiye; kbilge01@gmail.com

This study investigates microteaching in mathematics method courses, focusing on how mathematics teacher educators (MTEs) prepare primary teachers for teaching mathematics. We surveyed 65 MTEs across 71 Turkish universities, examining their approaches to topic selection, session settings, and assessment and feedback in microteaching. We found a preference for MTE-led topic selection and peer-based teaching sessions. Most MTEs tailor evaluation rubrics to individual needs, emphasizing a personalized assessment and feedback process. Furthermore, revealing evaluation criteria before sessions was noted to improve feedback transparency. The study underscores the impact of these practices on formative assessment and feedback.

Keywords: Microteaching, Mathematics methods courses, Mathematics teacher educators.

Introduction

Primary teachers are essential in developing foundational skills in learners during critical learning periods. Unlike specialized subject teachers, primary teachers cover a range of disciplines, including literacy, social studies, and mathematics. To prepare them, teacher training programs include method courses that focus not only on subject knowledge but also on effective teaching strategies and creating supportive learning environments (Strawhecker, 2005). Recognizing the importance of these courses, we conducted an in-depth study on the structure of mathematics method courses managed by MTEs. This paper presents our preliminary analysis and insights into microteaching within these courses from the MTEs' perspectives. We explore how these practices influence formative assessment and feedback, which are crucial for nurturing teacher candidates and achieving broad educational objectives of enhancing mathematical teaching skills (Bosica et al., 2021). This study addresses two main research questions: (1) Why do/ don't MTEs employ microteaching in method courses? and (2) How do MTEs structure the implementation of microteaching?

Literature review: microteaching in teacher preparation

Microteaching is a key component in teacher education, renowned for its effectiveness in improving teaching skills and easing the shift from theory to practice. It creates a simulated environment that enhances practical teaching abilities—like concept explanation, material preparation, and performance analysis—in small groups, as highlighted by Brown (1976). It also builds essential competencies for effective mathematics instruction (Dayal & Alpana, 2020; Cheng, 2017) and boosts self-efficacy and confidence in teacher trainees. Peker (2009) notes that microteaching significantly reduces teaching anxiety, improving confidence in lesson delivery. Batten (1979) and Majoni (2017) describe it as a method that focuses on specific skills, shortens teaching time, and reduces class size, effectively linking theoretical knowledge with actual teaching. Additionally, Elias (2018) stresses that microteaching facilitates feedback for behavioral change in candidates, adapting to individual needs and promoting a focused environment for skill development.

Microteaching's effectiveness largely stems from the tailored feedback it provides, which enhances teaching competencies. Brodsky and Doherty (2010) highlight feedback's crucial role in fostering learning and self-assessment, helping trainees identify their strengths and areas for improvement. Zhongji (2006) also notes that incorporating student feedback into microteaching significantly boosts teaching skills, making the training more efficient. Beyond skill development, feedback enhances motivation and interest among trainees, as noted by Özcan and Gerçek (2018). Peer feedback in microteaching can influence perceptions of 'good teaching', focusing on presentation and style, which affects evaluation processes (Vander Kloet & Chugh, 2012). Additionally, Mathew (2018) shows that feedback in microteaching creates a psychological environment conducive to improvement.

Building on these insights, Saraçoğlu and colleagues (2018) highlight that microteaching enables pre-service teachers to assess their own teaching proficiency in mathematics, providing a critical platform for self-critique and peer feedback. This reciprocal feedback process, as Semerci (2000) points out, not only enhances prospective teachers' self-evaluation skills but also improves their teaching performance through insights from peer evaluations. Learning from peers helps prospective teachers refine their teaching methods. These collective findings underscore microteaching's vital role in evaluating and enhancing teaching competencies, especially in mathematics, preparing teacher candidates for their future roles. Thus, microteaching is essential both as a feedback mechanism and a preparatory tool, contributing significantly to the candidates' professional readiness.

Methods

In this study, we share a portion of our broader research that explores six key areas in the design and structure of mathematics method courses within primary teacher education programs: the objectives, content delivery approaches, the design of teaching-learning situations, assessment and evaluation, collaborative endeavors, and research activities. To address each dimension, a mixed-methods survey comprising both open and closed-ended questions was crafted.

This study focuses on responses related to assessment and evaluation in microteaching, inspired by a literature review on diverse microteaching approaches (Mukuka & Alex, 2024). Three main characteristics of microteaching—planning, teaching in varied settings, and the evaluation-feedback cycle—were identified and reflected in the survey questions tailored to each aspect. Initially, the survey investigates whether MTEs include microteaching in their programs and their reasons for its inclusion or exclusion. Regarding planning, it queries how MTEs assign microteaching topics, from predetermined topics to allowing candidates to choose, including other possible methods. For the teaching settings, it asks about the intended audience, whether it is a real classroom, peers, a combination with video analysis, or other settings. Concerning evaluation, the survey explores if MTEs use a rubric for assessing microteaching and how these rubrics are sourced. It also examines the use of the evaluation tool, such as whether criteria are shared with candidates beforehand, if assessments are conducted solely by the MTE, or if peer evaluation is involved.

The participants of our study consisted of whole groups of MTEs who were responsible for teaching the mathematics method course within the primary teacher undergraduate programs at education faculties across Turkey. The target population consisted of 90 academics at 71 universities to whom the survey was sent via email. Responses were received from 73 MTEs, and out of these, 65 provided complete answers to all the questions in the survey. The remaining 8 indicated that they could not participate due to various reasons, such as being new to the course. Overall, the sample of 65 participants out of the population of 90 indicates a 72.22% representativeness.

The responses provided by MTEs in the survey underwent both qualitative and quantitative analyses. For the qualitative part, we focused on the open-ended responses, where reasons cited by participants were grouped under common themes to identify the underlying motivations for either implementing or refraining from microteaching in their courses. Quantitatively, the answers to the remaining survey items were scrutinized using descriptive statistics, with a particular emphasis on frequency counts. This analysis helped pinpoint prevailing trends regarding the selection of topics, the settings chosen for teaching, and the employed strategies for evaluation and feedback, providing a comprehensive overview of the current practices among MTEs.

Findings

Findings regarding the first research question

Findings related to the first research question revealed that out of 63 MTEs who responded to the open-ended question about the inclusion of microteaching practices in mathematics teaching courses, 2 left the questions unanswered. For those MTEs (n=50) who reported implementing microteaching, their objectives for providing feedback were categorized into six distinct themes. While several MTEs' responses spanned multiple themes, others were associated with a single theme exclusively.

Table 1: The reasons for implementing microteaching

Theme	n	Sample quotations
Putting theory into practices	27	*I have them do microteaching so that the theoretical knowledge they have learned can be implemented*
Improving trainee self-assessment	15	*It reduces their anxiety about giving a presentation. It provides instant feedback. Contributes to the development of evaluation and self-assessment skills*
Planning and preparations	10	*I think the best way to find answers to questions such as how to plan a lesson, how well they comply with this plan.*
Developing instructional skills	9	*I use it to provide an opportunity for teacher candidates to try their teaching skills, to see the mistakes and deficiencies they make during this time, and to correct them and try again.*
Developing subject matter knowledge	5	*Teacher candidates develop their own missing or additional subject knowledge by teaching.*
Developing classroom Management skills	4	*To gain awareness and experience in classroom management, to get to know students more closely and to have real classroom environment experiences.*

Examining Table 1 reveals that "putting theory into practice" is the predominant theme among the feedback topics MTEs address through microteaching for teacher candidates. This indicates a significant focus on the application of theoretical knowledge within practical teaching scenarios. Conversely, for the teacher educators who do not implement microteaching practices (n=13), the analysis yielded three distinct codes. These codes, which capture the reasons behind the absence of microteaching in their programs, are detailed in Table 2.

Table 2: The reasons for refraining from Microteaching

Codes	n	Sample quotations
Time constraints	6	*Incorporating practical applications into class time requires serious time.*
Excessive course load	3	*The most important reason is my high course load.*
Class size	4	*I think the classroom environment and class size are not suitable for a microteaching application*

Findings regarding the second research question

In this section, we present quantitative analysis regarding the second research question, more specifically, the methods MTEs use to deliver microteaching topics, the environments they prefer to conduct these sessions, the rubrics they use for assessment, and the evaluation and feedback approaches. The findings are presented in Table 3.

Table 3: MTEs' preferences to structure the microteaching

Category	Choices	n
MTEs' preferences of topic selection for microteaching	I determine the topics myself and distribute them to prospective teachers	28
	Prospective teachers decide on the topic of their choice	16
	Other	6
Preferred settings for microteaching implementations	Teaching in front of prospective teachers	34
	Teaching in real classroom environment and video recording and then showing it to prospective teachers	9
	Teaching in front of students in real classroom environment	7
Methods employed to create evaluation rubrics	I prepare it myself	26
	I adapt the existing	11
	I use ready-made	3
Preferences of MTEs in sharing evaluation criteria	I show the criteria to the prospective teachers in advance and evaluate them myself	27
	The prospective teacher is evaluated by his/her peers as well as myself.	16
	I evaluate only myself during the teaching without informing the criteria to the teacher candidates	7

With regard to choosing topics for microteaching sessions, a significant portion of MTEs (28 out of 50 respondents) opt to select the topics themselves to distribute to teacher candidates. In contrast, a smaller subset of 16 MTEs allows the candidates to choose their own topics.

As for the preferred settings for microteaching, a majority of the MTEs (34 out of 50) favor having teacher candidates conduct sessions in front of their peers. This method is more popular than teaching in a real classroom environment, which only garnered 7 responses, while 9 MTEs showed a preference for a combined approach where candidates teach in a real classroom and then review a video recording of their performance.

In terms of the assessment rubrics used for microteaching, most MTEs (26 out of 40) craft their own evaluation tools, indicating a trend towards tailor-made assessment strategies. Meanwhile, 11 MTEs adapt existing rubrics, and only 3 utilize pre-made rubrics without any alterations, underscoring the inclination towards personalized evaluation methods in microteaching.

With respect to the use of rubrics for evaluating microteaching sessions, the prevalent method among MTEs, with 27 indications, involves sharing the evaluation criteria with teacher candidates in advance, followed by the MTE conducting the assessment. A minority of 7 MTEs, however, choose not to reveal the criteria before the session and proceed with an independent evaluation. Additionally, 16 respondents incorporate peer feedback into the assessment process, combining it with their own evaluations.

Discussion

Discussion of research question one

Our study highlights the objectives behind MTEs' use of microteaching, primarily to transform theoretical knowledge into practical teaching application, bridging an important educational gap. This practice echoes Brodsky and Doherty’s (2010) emphasis on the importance of effective feedback for the development of teacher candidates, as it helps them identify strengths and improvement areas. MTEs value "Improving self-assessment," in line with Zhongji’s (2006) findings on the benefits of student feedback. Microteaching serves the dual purpose of honing teaching skills and fostering reflective practices, crucial for professional growth. MTEs also employ microteaching for comprehensive teacher preparation, including "Planning and preparations," "Developing instructional skills," "Development of subject matter knowledge," and "Classroom Management," recognizing the multifaceted nature of teaching and the diverse competencies required (Mathew, 2018). Drawing on insights from Saraçoğlu et al. (2018) and Semerci (2000), our findings argue that microteaching transcends a mere training approach, standing out as a critical, feedback-centric process that substantially contributes to teacher readiness. MTEs intentionally create a collaborative environment that bolsters peer learning and growth. Through microteaching, MTEs not only teach but also build a dynamic space for pre-service teachers to practice, reflect, and evolve, with feedback as a central element of this transformative experience.

Discussion of research question two

In this section, we discuss three aspects of microteaching—topic selection, preferred settings, and evaluation and feedback approaches.

Topic Selection

Our findings illuminate a pronounced inclination among MTEs to dictate microteaching topics, thereby exerting considerable influence over the trajectory of learning experiences and the nature of feedback provided to students. By selecting the topics, as we see it, MTEs are implicitly endorsing certain instructional priorities and competencies that they deem essential for prospective teachers to develop. This practice may, however, inadvertently narrow the scope of prospective teachers' potential growth. There is a potential risk that the feedback becomes tailored to a specific set of topics and attributes, potentially at the expense of a more holistic instructional approach that includes adaptability, responsiveness and pedagogical creativity. Moreover, Benton-Kupper (2001) underscores the value of microteaching as a scaffolded platform that enhances pre-service teachers' skills by offering them a space to engage in teaching practices and receive targeted feedback. The MTEs' involvement in topic selection is critical in this context, as it directs the areas of teaching that are emphasized and scrutinized. This deliberate guidance can have profound implications for the development of teaching skills, as it can ensure that feedback is specific, actionable, and aligned with the MTEs' vision.

Setting for microteaching

Our study indicates a significant inclination for conducting microteaching sessions in peer-based settings within a controlled environment, rather than directly within the real-world primary classroom context. Educators prefer these settings to facilitate a stable and secure environment where trainee teachers can develop their skills away from the complexities and unpredictability that come with an actual classroom (Benton-Kupper, 2001). This approach prioritizes an atmosphere conducive to formative feedback that is both immediate and specific, without the distractions and challenges that a typical primary classroom might impose. The benefit of such targeted and immediate feedback has been noted for its positive effects on the professional development of teacher candidates (Hidayah & Indriani, 2021). Nonetheless, it is recognized that feedback derived from a peer-based setting may inherently differ from that garnered in a real-classroom scenario. In a true classroom setting, feedback is not only prompt but also enriched by the real-life dynamics of classroom interaction, which is vital for fostering a teacher's ability to adapt and develop responsive teaching techniques (Sen, 2009). Traditional microteaching sessions in a university setting, with the oversight of MTEs, attempt to bridge the gap between educational theory and practice (Cheng, 2017), but still, they fall short of emulating the full spectrum of challenges in a primary classroom. Reflecting on this limitation, some researchers advocate for integrating microteaching sessions into actual primary classrooms to provide teacher candidates with a comprehensive and authentic teaching experience (Peker, 2009).

Evaluation and feedback approaches

Our findings show that most MTEs prefer to design or tailor their evaluation rubrics, highlighting the importance of context-specific, learner-centered feedback for improving student understanding and informing teacher insights on teaching effectiveness (Haug & Ødegaard, 2015). MTEs commonly share evaluation criteria with candidates beforehand, promoting transparency and structure in the feedback process. This approach supports Megawati's (2018) findings that peer assessments in microteaching enhance teaching skills by boosting confidence within a supportive learning atmosphere. Moreover, the use of feedback forms significantly influences teacher candidates' notions of effective teaching and may affect their self-concept and teaching approaches, as noted by Vander Kloet and Chugh (2012). In contrast, a smaller group of MTEs who choose not to disclose criteria prior to teaching might aim to encourage an independent assessment of teaching abilities, potentially leading to the development of more spontaneous teaching skills. This feedback strategy may lead candidates to display a wider array of teaching behaviors, contributing to a more comprehensive development, suggesting a need to refine evaluation methods to better support educational objectives.

Acknowledgement

This presentation stemmed from the first author's PhD studies.

References

Batten, H.D. (1979). *Factors Influencing the Effectiveness of Microteaching in a Teacher Education Programme*. Unpublished PhD thesis (Stirling, Department of Education, University of Stirling).

Benton-Kupper, J. (2001). The Microteaching Experience: Student Perspectives. Education 3-13, *121*(4), 830-835.

Bosica, J., Pyper, J. S., & MacGregor, S. (2021). Incorporating problem-based learning in a secondary school mathematics preservice teacher education course. *Teaching and Teacher Education*, *102*, 103335. http://dx.doi.org/10.1016/j.tate.2021.103335

Brodsky, D., & Doherty, E. G. (2010). Providing effective feedback. *NeoReviews*, *11*(3), 117-122.

Brown, G. (1976). Using Microteaching to Train New Lecturers. *Educational Media International*, *13*(3), 12-16.

Cheng, J. (2017). Learning to attend to precision: The impact of micro-teaching guided by expert secondary mathematics teachers on pre-service teachers' teaching practice. *ZDM*, *49*(1), 279-289.http://dx.doi.org/10.1007/s11858-017-0839-7

Dayal, H., & Alpana, R. (2020). Secondary pre-service teachers' reflections on their micro-teaching: Feedback and self-evaluation. *Waikato Journal of Education*, *25*(1), 73-83. http://dx.doi.org/10.15663/wje.v25i0.686

Elias, S. K. (2018). Pre-service teachers' approaches to the effectiveness of micro-teaching in teaching practice programs. *Open Journal of Social Sciences*, *6*(5), 205-224. https://doi.org/10.4236/jss.2018.65016

Haug, B. S., & Ødegaard, M. (2015). Formative Assessment and Teachers' Sensitivity to Student Responses. *International Journal of Science Education*, *37*(4), 629-654. https://doi.org/10.1080/02607476.2015.1080424

Hidayah, N., & Indriani, L. (2021). Real time feedback in English microteaching practice: A case study on online learning. *Metathesis: Journal of English Language, Literature, and Teaching*, *5*(2), 155-167. http://dx.doi.org/10.31002/metathesis.v5i2.4004

Majoni, C. (2017). Assessing the effectiveness of microteaching during teacher preparation. *European Journal of Research and Reflection in Educational Sciences*, *5*(2).

Mathew, L. K. (2018). Comparative Study of Responses from Different Groups in Microteaching. *Journal of Medical Science and Clinical Research*, *6*(2), 806-809. https://dx.doi.org/10.18535/jmscr/v6i2.123

Megawati, F. (2018). Peer observation of teaching: Pre-Service Teachers' Perspectives for Better Performance. *Advances in Social Science, Education and Humanities Research, 125*(1), 124-127. http://dx.doi.org/10.2991/icigr-17.2018.30

Mukuka, A., & Alex, J. K. (2024). Review of research on microteaching in mathematics teacher education: Promises and challenges. *Eurasia Journal of Mathematics, Science and Technology Education*, *20*(1), 1-15. http://dx.doi.org/10.29333/ejmste/13941

Özcan, Ö., & Gerçek, C. (2018). Multidimensional analyzing of the microteaching applications in teacher education via videograph. *European Journal of Teacher Education*, *42*(1), 82-97. https://doi.org/10.1080/02619768.2018.1546285

Peker, M. (2009). The use of expanded microteaching for reducing pre-service teachers' teaching anxiety about mathematics. *Scientific Research and Essays,* *4*(9), 872-880. https://doi.org/10.12973/ejmste/75284

Saraçoğlu, G., Gürışık, A., & Furat, D. (2018). Opinions of english teacher candidates regarding the criticism made after micro-teaching applications. *Turkish Journal of Educational Sciences*, *16*(1), 58-76.

Semerci, N. (2000). The Effect of Critical Thinking on Developing Criticism Skills in Micro Teaching Course (FÜ Technical Education Faculty Sample). *Education and Science*, *25*(117).

Sen, A. I. (2009). A study on the effectiveness of peer microteaching in a teacher education program. *Education and Science*, *34*(151), 165-174.

Strawhecker, J. (2005). Preparing elementary teachers to teach mathematics: How field experiences impact pedagogical content knowledge. *Issues in the Undergraduate Mathematics Preparation of School Teachers: The Journal*, *4*(Curriculum),1-12.

Vander Kloet, M. A., & Chugh, B. P. (2012). An interdisciplinary analysis of microteaching evaluation forms: How peer feedback forms shape what constitutes "good teaching". *Educational Research and Evaluation*, *18*(6), 597-612.

Zhongji, H. (2006). Microteaching-Researching: an effective approach to improving the teaching skills of college teachers. *Journal of Lanzhou Institute of Education*, 3, 40-42.

Certainty-based marking as feedback in the context of formative assessment in mathematics lessons at school

Joerg Zender[1] and Martina Geisen[2]

[1]Private School, Frankfurt am Main, Germany; joerg@zender.xyz

[2]University of Potsdam, Germany; martina.geisen@uni-potsdam.de

Self-reflection is an essential skill needed to adjust one's learning process. However, training is necessary to develop this skill. Certainty-based marking is an easy-to-integrate method of formative assessment that strengthens predictive accuracy about one's answers to mathematics problems. Knowing better about one's knowledge might lead to better self-reflection. An explorative study with German 5th graders is presented here, in which the influence of certainty-based marking in formative assessment on pupils' mathematical self-reflection during mathematics lessons is examined. Initial findings show that the certainty and the accuracy with which the pupils state their knowledge rise over time.

Keywords: Formative assessment, certainty-based marking, self-reflection, middle school, feedback.

Introduction

Self-reflection is a cognitive process involving conscious observation of one's thoughts, feelings, and actions and is essential for understanding individual learning, e.g. for mathematics lessons at school. In dealing with heterogeneity, various forms of differentiation are used in mathematics lessons in which, among other methods, learners choose from the options provided (e. g. Hußmann & Prediger, 2007). However, this can be a challenge for students. They must first learn to assess themselves correctly and then make a selection based on this to control their learning and take responsibility for it in the long term. So, promoting students' self-reflection can improve teaching and learning outcomes. One possibility for training students' self-reflection could be the method of certainty-based marking (CBM) in the context of formative assessment. CBM is a testing format that requires students to express a degree of certainty in their responses, which will be considered for grading.

This article presents the initial findings of an ongoing exploratory study that investigates how CBM in formative assessment influences the self-reflection of 5th-grade learners regarding mathematics at school. In the following sections, CBM is first explained to address its potential in the context of formative assessment and the need for research in this area. Then, the methodological approach is described, initial results are presented, and finally, the results are reflected upon.

Certainty-based marking as feedback in the context of formative assessment

Studies on feedback in the context of formative assessment were carried out by Brensing et al. (2021). The results show that combined formative assessment and feedback can be a powerful tool for fostering mathematical learning. In their study on a first-semester course of mathematics for engineering, counselling after a formative assessment leads to better results, and even weak students with low mathematical knowledge could come close to the mathematical performance of the middle group. The feedback was given in individual counselling sessions, which was quite work-intensive and time-consuming, but it helped the students to adjust their learning and perform better. Following Kruger and Dunning (1999), in these counselling sessions it was revealed that the weak students did not consider themselves weak. But they thought they performed much better than they did. They first needed a realistic self-reflection to help them understand what they did not understand to start an efficient learning process.

CBM is an assessment approach developed and first used by Gardner-Medwin (1995) for teaching physiology, medicine, and maths at University College London. The genesis of CBM can be traced back to the desire for a more nuanced and comprehensive assessment methodology that goes beyond the limitations of traditional grading systems (ibid.). Traditional grading systems often provide a single numerical score for a student's solution to a problem, leaving little room for insight into the students' thought process or the level of confidence in their answer (ibid.). So, instead of assigning a fixed score to a solution, CBM allows students to express their confidence in their answers. Students typically provide the solution to a problem and add a confidence rating, such as high, medium, or low, influencing the grading (e. g. see Table 1). Certainty-based marking encourages students to think critically about their responses and reflect on their understanding of the material. This approach recognizes that learning is a dynamic process, and students may possess varying degrees of confidence even when their answers are correct. By incorporating certainty levels into assessment, educators gain insights into students' self-awareness and understanding of the subject matter.

Later, Yuen-Reed and Reed (2015) simplified the confidence rating to two confidence levels and grades without malus points, in contrast to Gardner-Medwin (1995). The two options of "unsure" and "very unsure" are combined into just one option of “unsure” because Yuen-Reed and Reed (2015) argue that this distinction is very subjective.

Today, studies on various facets of CBM are available. It has been used in assessment but more in self-assessment and summative assessments (Gardner-Medwin, 2019). In Germany, CBM was, for example, conducted at the RheinMain University of Applied Sciences in a mathematics course with 43 third-semester students in the Department of Engineering (Kanzinger & Gehrig, 2022). Four tests were conducted, each with five subject-specific single-choice questions and CBM. Results were sent to the students at the end of the day the tests had been conducted. An online survey among the students as an evaluation of CBM took place at the end of the semester. Results indicate that CBM could promote the students' ability to self-reflect on tests during the semester. Almost all students stated that CBM helps them to reflect on their learning. There was no control group in this study (ibid.). Following these ideas, integrating formative assessment with CBM could represent a powerful and nuanced approach to understanding and enhancing student learning, particularly in disciplines like mathematics. It has been tried in medicine (Hendriks et al., 2019) but only in one assessment inside a series of assessments and in veterinary (Valero & Cárdenas, 2017). Foster (2016) used CBM in mathematics in school (and called it Confidence Assessment, CA) and investigated how pupils deal with this approach and respond. For this purpose, a ten-task questionnaire on negative numbers and a certainty grading with ten steps was carried out with different classes from age 10 to age 13. Based on this test, Foster (2016) stated that the pupils were well-calibrated but demanded to repeat the test since they then understood their certainty better. According to Foster (2016), research is needed into what effects will be there if CBM is done more often. A further study by Foster (2022) showed mixed results regarding the effects of summative assessment. Conducted in four schools with n=475 pupils, but varying in time (from 3 weeks to an entire school year), Foster found that in schools pupils became better than the control group, and in others, they did not.

Formative assessment takes place parallel to the learning unit or course to support the learning process and to improve individual learning (cf. Geisen & Zender, 2023; see also Brookhart, 2010; Cizek, 2010; Maier, 2010; Gikandi et al., 2011) and enables teachers to adapt learning opportunities to the needs of the respective learners (cf. Black & Wiliam, 2009). Introducing CBM within formative assessment could take this process a step further. Students could receive feedback on their answers and be prompted to self-reflect and express their confidence levels in their answers, for which they also receive feedback. This dual-layered evaluation could acknowledge the dynamic nature of learning and encourage metacognition – the ability to think about one's thinking. The integration of formative assessment with CBM could cultivate a learning environment that is responsive, adaptive, and focused on continuous improvement. On the one hand, students could become active participants in their learning journey. On the other hand, educators could receive valuable data, not only one-dimensional, on how much the students did correct, but also a second dimension, which is how certain they are. This leads to new insights if a topic remains unclear for the students, although they did perform well. Alternatively, students could be very confident in a topic but answer the questions wrongly, which can uncover misconceptions. Based on the data, educators can tailor their instructional strategies to the different types in this now two-dimensional outcome.

The literature shows that CBM is used in the university context (and mainly in medicine) but not much in school so far. It is also not commonly known or used in Germany. Therefore, the study presented below can be seen as a starting point for using CBM in the context of formative assessment in school mathematics lessons.

A weekly CBM test in mathematics lessons - an exploratory study in fifth grade

Objectives

CBM is rare and plays a minor role in assessment. Nevertheless, the wide range of Gardner-Medwin’s publications (e. g. 2019) show a promising use of CBM in self-assessments. However, there is a need for research concerning CBM in the context of formative assessment. Furthermore, self-reflection should already be used at school. Using CBM in formative assessment could train pupils at school to reflect on their knowledge and improve on this reflection, a necessary skill for individual learning. Therefore, self-reflection should be initiated early in a learner's biography. However, there is still a need for research in this area. An exploratory study is therefore being conducted to investigate the influence of CBM in formative assessment on middle school pupils’ mathematical self-reflection. Instead of directly measuring self-reflection, the accuracy of the confidence in the answers was measured by how often a student was correct when sure about the answer or how often a student was not correct when unsure about the answer.

Method and sample

To measure the influence of CBM, an experimental fifth-grade class was chosen from a full-time German grammar school. The class had n=15 pupils aged 10 and 11 (six male and nine female). The pupils came from the same elementary school, and their respective performance levels in math lessons were heterogeneous. They had six hours of mathematics lessons each week, in which the last lesson of each regular week (excluding vacations, weeks in which there were reports and in which a class test was written) was used for a short test on the content of the previous lessons in that week. By the time this article was written, sixteen tests had been set. The tests were conducted weekly and contained eight open questions (no multiple choice). Additionally, they had a certainty question after each mathematical problem about how sure the students were about their solution. There were two checkboxes: “I am □ sure □ not sure.”. Based on the concept of Yuen-Reed and Reed (2015), the certainty levels were reduced from three to two to make it simpler for the pupils because of their age. The rating was 2 points for a correct and sure answer, -2 points if the answer was wrong but sure, 1 point for a right but unsure answer, and 0 points if the answer was wrong and unsure. If no checkbox was used, 0 points were assigned regardless of whether the answer was correct or wrong (see Table 1). So, it is always better to state at least to be unsure than not using the checkbox. Consequently, the pupils have used the checkboxes. Also, in this system, it is easy to understand that it is better to be honest about the certainty of your answer.

Table 1: CBM points scheme (based on Yuen-Reed & Reed, 2015)

	sure	unsure	nothing
correct	2	1	0
incorrect	-2	0	0

Empirical analysis and initial findings

Regarding accuracy, the class test results were investigated in two ways: first, by how many correct answers the pupils gave and whether they were sure about this, and second, by how many incorrect answers the pupils gave and whether they were unsure about this. The results are displayed in Table 2.

Table 2: Pupils’ prospective validity in percent

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8
sure and correct	67%	70%	66%	59%	83%	86%	75%	87%
unsure and incorrect	52%	63%	68%	66%	85%	67%	80%	64%
mean	59%	67%	67%	62%	84%	77%	78%	76%

	Test 9	Test 10	Test 11	Test 12	Test 13	Test 14	Test 15	Test 16
sure and correct	88%	83%	89%	79%	76%	84%	84%	83%
unsure and incorrect	57%	77%	61%	49%	58%	69%	52%	58%
mean	72%	80%	75%	64%	67%	77%	68%	71%

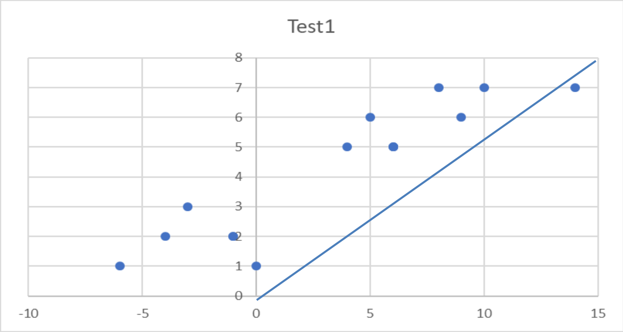
The first four tests looked almost the same, but something changed, and the pupils started to rank higher. The diagrams in Figure 1 (see next page and page after next) show each pupil represented by one dot. The dot is positioned on the x-axis based on how many points the pupil got in the test and on the y-axis based on how many correct answers were given by this pupil. A perfect accuracy would mean the pupil is on a straight line with a gradient of ½. The diagrams show that the points move towards that line over time, especially after the fourth test.

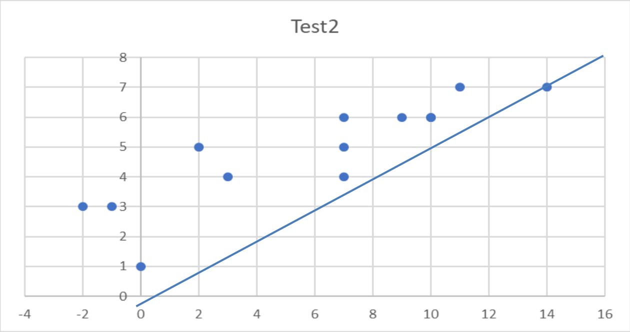
Conclusions and Discussion

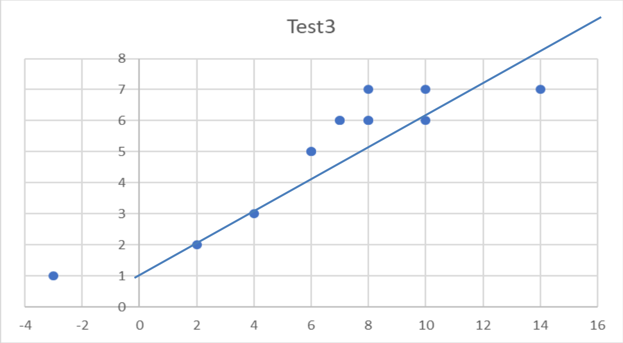
This paper focuses on using CBM within formative assessment in school mathematics lessons, a strategy that could enhance the prospective validity of pupils' answers to mathematical problems. Understanding one's expertise can lead to improved self-reflection, a crucial skill for adjusting one's learning process. Therefore, it is necessary to investigate the influence of CBM in formative assessment on the mathematical self-reflection of pupils at school. Furthermore, research is needed into the potential effects of more frequent CBM testing (Foster, 2016). To address these questions, we conducted an empirical study, testing a class of 5th graders at a German grammar school with CBM tests almost every week for a school year. At the time of writing this article, we had administered sixteen tests. After the first four tests with nearly the same prospective validity, the rate suddenly went up. Pupils got better at telling what they had learned and what not. As stated in the methods section, the self-reflection itself was not measured, but a better accuracy could speak in favour of a better self-reflection. At least, one must know about one's knowledge as a basis of self-reflection. If the pupils used their knowledge to learn and practice precisely what they had identified as their weaknesses before, this would be a step towards self-regulated learning. However, there has been no further investigation into the consequences of the test results for the pupils so far.

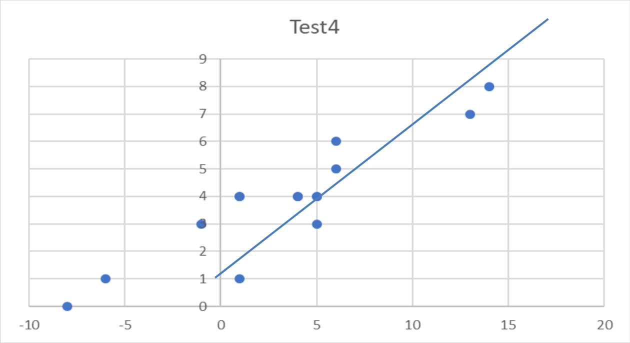
While the category “sure and correct” remains on a high level, the opposite, “unsure and incorrect” seems not to benefit from the CBM treatment. One possible conclusion is that the repeated use of CBM manages to reduce overconfident pupils, but on the other hand, it may create some insecurities. Regarding the origin of CBM, this effect is desired in the case of medical doctors but it may not be so much in favour when it comes to pupils in school. Open questions are whether this effect mainly concerns weaker pupils, who are generally more insecure about their mathematical knowledge.

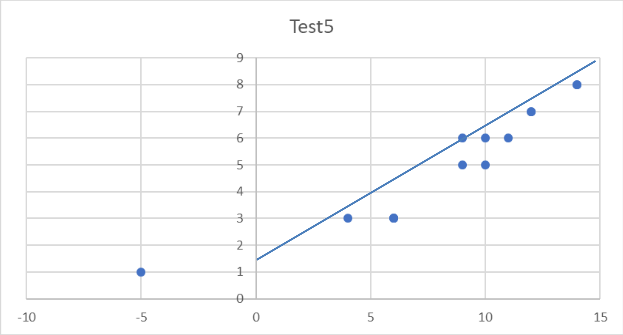
The presented exploratory study is a starting point and is still ongoing, so the findings must be interpreted cautiously. In addition, these findings relate exclusively to one experimental group, so to a small sample. Therefore, the study will initially be continued in the current groups and then expanded to include further groups so that future results will be based on a broader database.

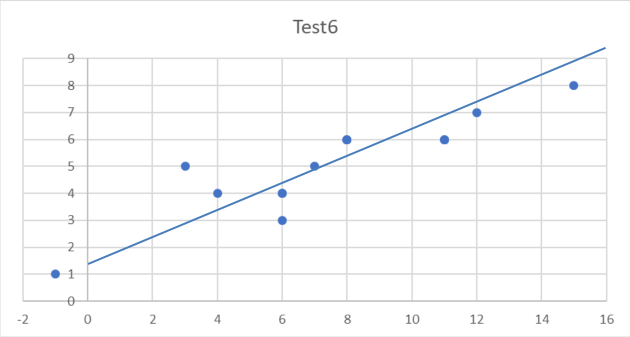

Test1
0 1 2 3 4 5 6 7 8
-10 -5 0 5 10 15

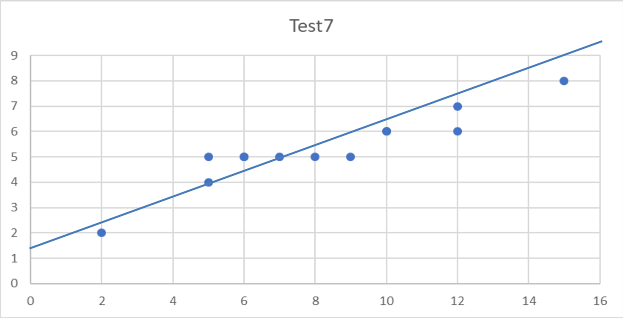

Test2
0 1 2 3 4 5 6 7 8
-4 -2 0 2 4 6 8 10 12 14 16

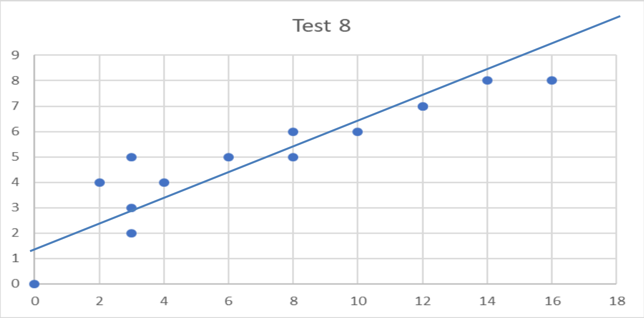

Test3
0 1 2 3 4 5 6 7 8
-4 -2 0 2 4 6 8 10 12 14 16

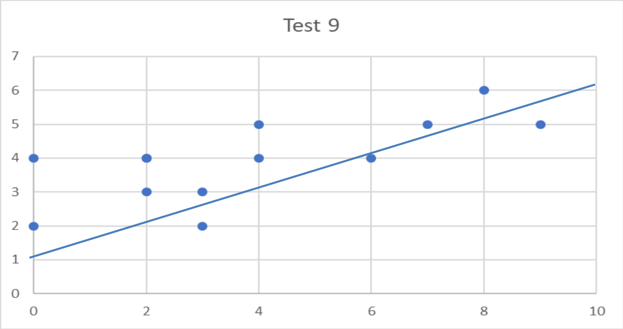

Test4
0 1 2 3 4 5 6 7 8 9
-10 -5 0 5 10 15 20

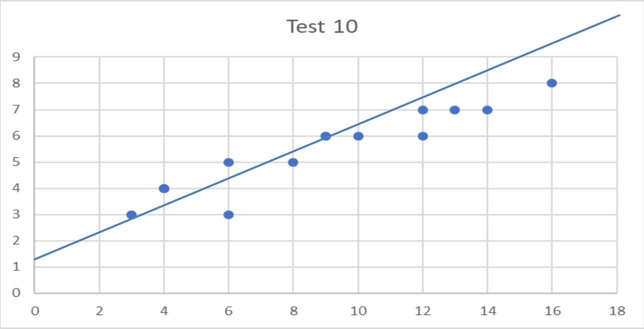

Test5
0 1 2 3 4 5 6 7 8 9
-10 -5 0 5 10 15


Test6
0 1 2 3 4 5 6 7 8 9
-2 0 2 4 6 8 10 12 14 16


Test7
0 1 2 3 4 5 6 7 8 9
0 2 4 6 8 10 12 14 16


Test 8
0 1 2 3 4 5 6 7 8 9
0 2 4 6 8 10 12 14 16 18


Test 9
0 1 2 3 4 5 6 7
0 2 4 6 8 10


Test 10
0 1 2 3 4 5 6 7 8 9
0 2 4 6 8 10 12 14 16 18

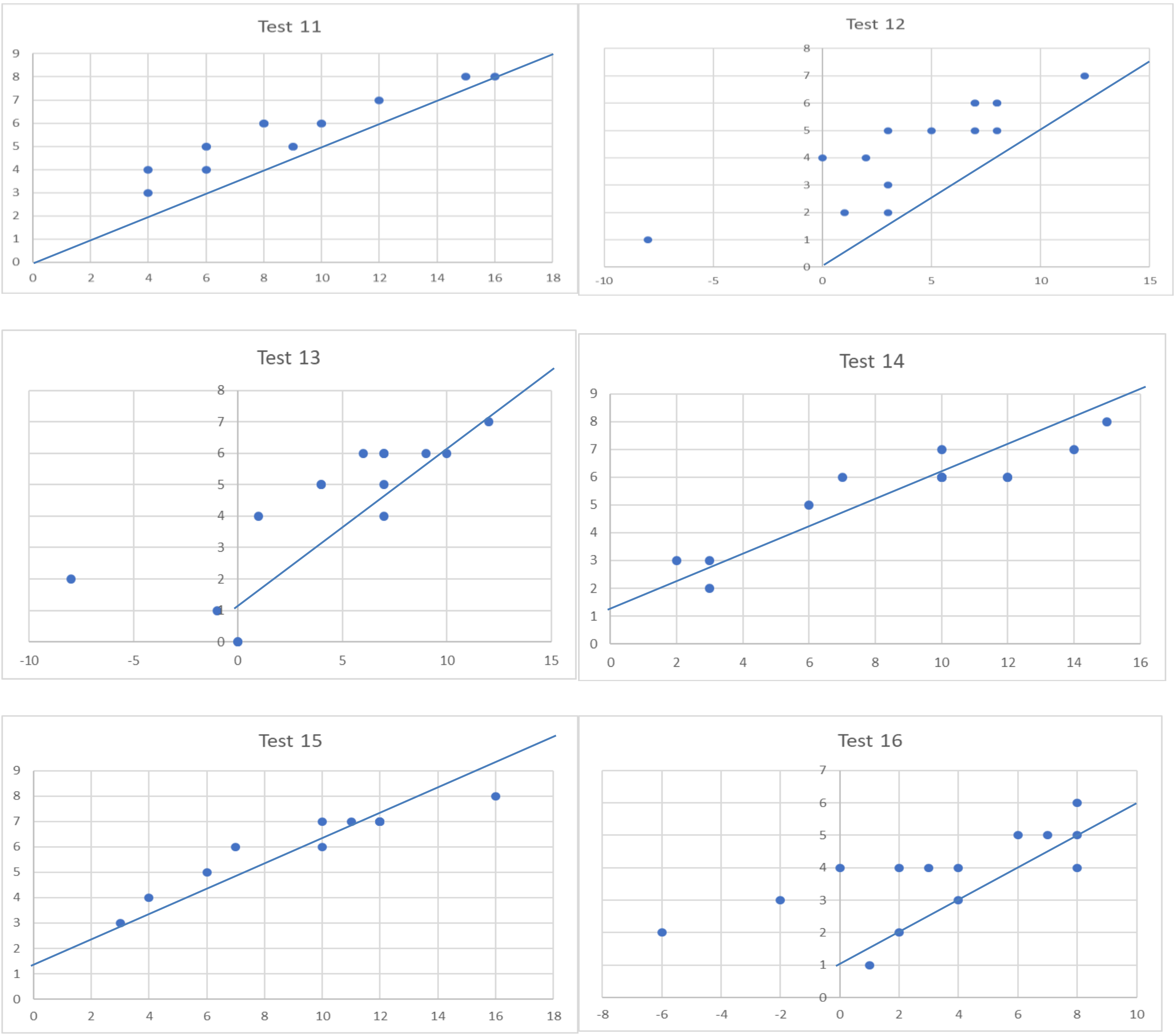


Figure 1: Test results of the CBM class (Each point represents the results of one pupil. The x-axis is the final score, while the y-axis is the number of correct answers.)

References

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*, 5–31.

Brensing, M., Dannewald, T., Kanzinger, A., Mayer, U., & Zender, J. (2021). Counselling in the introductory phase of studies. *Zeitschrift für Hochschulentwicklung, 16*(1), 117–136.

Brookhart, S. M. (2010). *Formative assessment strategies for every classroom: An ASCD action tool.* Association for Supervision and Curriculum Development.

Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. In Andrade, H., Cizek, G. J. (Eds.), *Handbook of formative assessment* (pp. 3–17). Routledge. https://doi.org/10.4324/9780203874851

Foster, C. (2016). Confidence and competence with mathematical procedures. *Educational Studies in Mathematics, 91*(2), 271–288. https://doi.org/10.1007/s10649-015-9660-9

Foster, C. (2022). Implementing confidence assessment in low-stakes, formative mathematics assessments. *International Journal of Science and Mathematics Education, 20*(7), 1411-1429.

Gardner-Medwin, A. R. (1995). Confidence assessment in the teaching of basic science. *ALT-J*, *3*(1), 80-85.

Gardner-Medwin, T. (2019). Certainty-based marking. In Clegg, K. (Ed.), *Innovative Assessment in Higher Education: A Handbook for Academic Practitioners*, 141–150.

Geisen, M. & Zender, J. (2023). Formative Assessment in online courses – Ideas and Experiences. In M. Ludwig, S. Barlovits, A. Caldeira, & A. Moura (Eds.), Research On STEM Education in the Digital Age. *Proceedings of ROSEDA 2023: Research Online STEM Education in the Digital Age* (pp. 25-32). WTM-Verlag.

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & education*, *57*(4), 2333–2351.

Hendriks, W. J. A. J., Bakker, N., Pluk, H., de Brouwer, A., Wieringa, B., Cambi, A., Zegers, M., Wansink, D. G., Leunissen, R., & Klaren, P. H. M. (2019). Certainty-based marking in a formative assessment improves student course appreciation but not summative examination scores. *BMC Med Educ* 19, 178. https://doi.org/10.1186/s12909-019-1610-2

Hußmann, S., & Prediger, S. (2007). Mit Unterschieden rechnen - Differenzieren und Individualisieren. *Praxis der Mathematik in der Schule*, *49*(17), 2–8.

Kanzinger, A., & Gehrig, E. (2022). Certainty-based-marking - eine kompetenzorientierte Prüfungsmethode. In IDMI-Primar Goethe-Universität Frankfurt (Eds.), *Beiträge zum Mathematikunterricht 2022.* 56. Jahrestagung der Didaktik der Mathematik. WTM-Verlag für wissenschaftliche Texte und Medien. http://dx.doi.org/10.17877/DE290R-23377

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, *77*(6), 1121–1134.

Maier, U. (2010). Formative assessment—A promising concept for improving instruction and classroom assessment. *Zeitschrift für Erziehungswissenschaft*, *13*, 293–308.

Valero, G., & Cárdenas, P. (2017) Formative and Summative Assessment in Veterinary Pathology and Other Courses at a Mexican Veterinary College. *Journal of Veterinary Medical Education, 44* (2), 331–337. https://doi.org/10.3138/jvme.1015-169R

Yuen-Reed, G., & Reed, K. B. (2015). Engineering Student Self-Assessment through Confidence-Based Scoring. *Advances in Engineering Education*, *4*(4), 1–23

Posters

Assessment in the mathematics classroom in relation to how students are in(ex)cluded in mathematics

Maria Silwer

Malmö University, Malmö, Sweden; maria.silwer@mau.se

Keywords: Classroom assessment, students' perspectives, teachers' perspectives, mathematical reasoning.

Introduction and background

In this poster I would like to discuss the design of my PhD-project on assessment in the mathematics classroom in relation to how students are in(ex)cluded in mathematics, in early school years (ages 10-12), in Sweden. This poster uses in(ex)clusion as a concept for inclusion and exclusion. Here inclusion in mathematics is when students can access and participate in mathematics learning (Roos, 2019b, referred to in Roos, 2023). Exclusion is always present in the process, when working for inclusion (Valero, 2021, referred to in Roos, 2023).

This poster uses the following definition of assessment practice: All decisions the teacher makes when assessing, e.g., what is assessed, which tasks to use, when and in which situations (Boistrup, 2017). Hence, assessment is a part of learning and education, and therefore also in mathematics education. In mathematics classroom assessment the teacher gathers information about students' mathematical knowing to support learning but also to improve teaching practice (Nieminen et al., 2023). Assessment affects how the students are positioned when learning mathematics due to roles and responsibilities (Nieminen et al., 2023). This could indicate that students meet different kinds of assessment practices in different classrooms, which in turn in(ex)clude the students in relation to learning mathematics. Hence, this poster assumes that assessment is present in all teaching and is shown to the student when given as feedback (Boistrup, 2017).

Plan for upcoming study

The tentative aim of this study is to build an understanding of assessment in the mathematics classroom in relation to how students are in(ex)cluded in mathematics. The mathematical focus will be reasoning due to two reasons. One, different kinds of reasoning are generated by problem solving (Boesen et al., 2010, referred to in Säfström et al., 2024), which in turn is shown to be beneficial for students' learning of mathematics (Liester & Cai, 2016, referred to in Säfström et al., 2024). Two, reasoning can exist regardless of mathematical content, e.g. arithmetic or geometry. Hence, this study will explore students' perspectives on assessment in mathematics and how students are in(ex)cluded in relation to learning mathematics through reasoning in the classroom. The teacher matters in the classroom (Terhart, 2011) and in the assessment, so this study will also focus on the teacher. By doing this, it will be possible to gain insights into how equity-driven assessment practices could be used in the mathematics classroom. To meet the aims of this study, its design will be developed systematically structured by the following parts.

First, the aim is to examine how students are in(ex)cluded in relation to learning mathematics in the classroom. The tentative research question is: What traces of in(ex)clusion can be identified in interviews with students and teachers? To answer this question interviews will be conducted with students and teachers, with a focus on their perspectives of assessment in the mathematics classroom. The results will be presented in the first article.

Second, the aim is to compare responses from students and teachers, to identify potential differences in assessment practices in relation to in(ex)clusion, within and between classrooms. The tentative research questions are: What similarities and differences due to in(ex)clusion can be identified in students' and teachers' perspectives on assessment in mathematics? What similarities and differences regarding assessment practices within and between classrooms can be identified? The earlier interviews will be used as a pilot study to design a survey with multiple-choice questions, to further examine students' and teachers' perspectives on assessment in the mathematics classroom. The results will be presented in the second article.

Third, the aim is to examine assessment practices in relation to how students are in(ex)cluded in relation to reasoning in mathematics education. The tentative research question is: How do different assessment practices in(ex)clude students in relation to reasoning in mathematics? A survey with open-ended questions will be used to examine teachers' perspectives on their own assessment practice. The results will be presented in the third article.

Fourth, the aim is to find ways to assess students reasoning in mathematics, on a classroom level, where the assessment practice promotes equity for all students. The tentative research question is: What can an assessment practice that promotes equity for all students look like? This case study will draw on findings from the earlier partial studies to develop assessment practices together with a teacher, in the mathematics classroom. The results will be presented in the fourth article.

References

Boistrup, L. B. (2017). Assessment in mathematics education: A gatekeeping dispositive. In H. Straehler-Pohl, N. Bohlmann, & A. Pais (Eds.), *Disorder of Mathematics Education: Challenging the Sociopolitical Dimensions of Research* (pp. 209–230). Switzerland: Springer International Publishing.

Nieminen, J. H., Bagger, A., Padilla, A., & Tan, P. (2023). Student Positioning in Mathematics Assessment Research: A Critical Review. *Journal for Research in Mathematics Education*, *54*(5), 317–341. https://doi.org/10.5951/jresematheduc-2022-0030

Roos, H. (2023). Students' voices of inclusion in mathematics education. *Educational Studies in Mathematics*, *113*(2), 229–249. https://doi.org/10.1007/s10649-023-10213-4

Säfström, A. I., Lithner, J., Palm, T., Palmberg, B., Sidenvall, J., Andersson, C., Boström, E., & Granberg, C. (2024). Developing a diagnostic framework for primary and secondary students' reasoning difficulties during mathematical problem solving. *Educational Studies in Mathematics, 115*(2), 125–149. https://doi.org/10.1007/s10649-023-10278-1

Terhart, E. (2011). Has John Hattie really found the holy grail of research on teaching? An extended review of Visible Learning. *Journal of Curriculum Studies*, *43*(3), 425–438. https://doi.org/10.1080/00220272.2011.576774

The importance of formative assessment in developing student teachers' teaching practice

Daniel de Oliveira Lima[1]

[1]University of the State of Rio de Janeiro, Cap-UERJ, Department of Mathematics and Drawing, Rio de Janeiro, Brazil; danielprof2006@gmail.com

Keywords: Mathematics assessment, pedagogical practices, teacher training, formative assessment, teacher development.

Introduction

Assessment is essential in teacher education as it elucidates how and what enhances students' school performance in class. For instance, Lima (2022) delves into the significance of teacher training in comprehending assessment in mathematics. At the University of State of Rio de Janeiro (UERJ), student teachers learn about pedagogical practices in mathematics during their mathematics undergraduate studies. This course's curriculum emphasizes preparing students for the challenges they will encounter in their future school classes. The course content integrates concepts from primary education with a focus on active methodologies and formative assessment. Ideas from Anderson and Palm (2018), Buchholtz, et al. (2018) and Lima (2022) inspire reflective discussions regarding formative assessment. This article aims to explore the role of formative assessment in the course and whether having a focus on formative assessment in an undergraduate mathematics can benefit student teachers in their teaching practice.

Theoretical framework

According to Anderson and Palm (2018), assessment is a process that involves teachers and their students in collecting, analyzing, interpreting, discussing, and utilizing information regarding students' learning. Assessment fulfills two roles: summative and formative. The former involves assessing school students' performance, allowing teachers to rank students and provide an overview of their learning progress. The latter concentrates on the learning process and emphasizes feedback and offers detailed personalized feedback.

Lima (2022) suggests that formative assessments aim to enhance learning, essentially making them integral to the teaching process. Formative assessment, when conducted solely by the teacher, becomes an intrinsic part of teaching. It's important to note that since teaching methods vary, so do assessment methods. It is necessary to highlight that there are different ways of teaching, so there are other ways of evaluating. Buchholtz, et al. (2018) argue that any assessment can be either formative or summative, contingent upon the teacher's perspectives.

Lima (2022) distinguishes between summative and formative feedback. Summative feedback centers on grades and occurs at the culmination of a process, evaluating a student's overall performance. In contrast, formative feedback transpires throughout the process, pinpointing errors and devising strategies to address weaknesses. Moreover, it facilitates students' self-regulation and fosters autonomy.

Research questions and aim

Lima (2022) argues that mathematics undergraduate programs must incorporate assessments in mathematics. Consequently, as the author and professor specializing in pedagogical practices in mathematics, he introduced this subject in his class. Building upon this premise and drawing from the research of Lima (2022) and Buchholtz, et al. (2018) the present study aims to explore the research question: How has a focus on formative assessment in their undergraduate mathematics degree benefited student teachers in their teaching practice?

Analyzing some answers from students

To address the research question, two student reports about student experience in the course were analyzed. Out of eight students, these were the ones who participated the most in class and showed higher levels of engagement. The reports were analyzed through qualitative research using document analysis.

Student A expressed, "In this specific subject, I performed well; I learned extensively and gained insights into classroom dynamics, student challenges, and effective strategies for addressing individual learning processes”. This student emphasized the importance of recognizing and addressing student difficulties, which is often overlooked, particularly among math educators.

Student B shared, "The course significantly enhanced my ability to create inclusive and engaging learning environments for my students. I've noticed a substantial increase in their participation and enthusiasm”. Despite acknowledging that more consistent attendance could have benefited her learning, she found the course instrumental in improving her teaching practices.

Conclusions

Throughout the course, students were challenged to create lesson plans and activities, apply active methodologies, and receive specific feedback. Additionally, the class examined the significance of teacher development, encompassing knowledge, professional culture, pedagogical tact, teamwork, and social commitment. It is evident that the course had a positive impact on students' preparations for their future careers as educators.

References

Andersson, C. & Palm, T. (2018). Reasons for teachers’ successful development of a formative assessment practice through professional development – a motivation perspective. *Assessment in Education: Principles, Policy & Practice, 25*(6), 576–597. https://doi.org/10.1080/0969594X.2018.1430685

Buchholtz, N. F., Krosanke, N., Orschulik, A. B., & Vorholter, K. (2018). Combining and integrating formative and summative assessment in mathematics teacher education. *ZDM Mathematics Education, 50*, 715–72. https://doi.org/10.1007/s11858-018-0948-y

Lima, D. (2022). *Concepções de professores de matemática sobre a avaliação escolar: o caso da Escola Sesc de ensino médio* [Doctoral dissertation, Federal University of Rio de Janeiro]. Theses and Dissertations Archive. https://pemat.im.ufrj.br/index.php/pt/producao-cientifica/teses/104-2022/379-concepcoes-de-professores-de-matematica-sobre-avaliacao-escolar-o-caso-da-escola-sess-de-ensino-medio.

Perceptions of Effective Formative Feedback: A comparative Analysis Between Undergraduate Students and Mathematics Lecturers

Samah Taha, George Kinnear and Paola Iannone

University of Edinburgh, School of Mathematics; Samah.taha@ed.ac.uk ; G.kinnear@ed.ac.uk and Paola.innone@ed.ac.uk

Keywords: Formative feedback, feedback framework, comparative judgement.

In the context of formative e-assessment in mathematics, students often demonstrate misconceptions, which underlie their mistakes. This prompts an exploration of the role of feedback as cornerstone of formative e-assessment. Given the increasing reliance on technology in education and the need for adaptable assessment methods to meet the evolving needs of learners in the digital age, focusing on e-assessment becomes imperative (Evans, 2013). What feedback should mathematics lecturers provide within e-assessment to effectively support student learning? This poster presents the perspective of both undergraduate students and mathematics lecturers to the feedback provided by fourteen mathematics lecturers in response to a fictional student response to a formative e-assessment task.

High-quality feedback may influence student achievement (Mulliner & Tucker, 2017). However, despite substantial time and effort invested in generating feedback, there is a noticeable lack of research in higher education regarding its effectiveness (Price et al., 2010). Despite evidence of students' desire for feedback from their lecturers, the National Students Survey in the UK has documented student dissatisfaction since 2005, with students expressing discontent over the nature and timing of feedback (Price et al., 2010). In contrast, academics often believe their students are receiving timely, extensive, and informative feedback (Mulliner & Tucker, 2017).

Assessment feedback includes all feedback exchanges generated within assessment design, occurring within and beyond the immediate learning context, being overt or covert (actively and/or passively sought and/or received), and importantly, drawing from a range of sources (Evans, 2013). A comprehensive review by Lipnevich and Panadero (2021) presents fourteen models, complete with accompanying diagrams, explaining how feedback operates and identifying variables that may contribute to student engagement with it. Within this complex landscape of factors and interactions, our study draws a narrow focus on the content of the feedback message that mathematics lecturers provide to undergraduate students on formative e-assessment tasks. In this poster, we address two questions:

RQ1: Do mathematics lecturers generally agree on the characteristics of the effectiveness of formative feedback?

RQ2: Do the lecturers' perceptions of effective feedback align with those of their students?

Building upon the research conducted by Evans et al. (2022), which highlights the efficacy of comparative judgment in studying the quality of mathematical explanations, we adopt this approach to analyse the quality of feedback given on e-tasks. Comparative judgment is a widely utilised method in educational research, employed to evaluate student's essays, laboratory reports, and abstract constructs such as conceptual understanding, problem-solving, and mathematicians' proof conceptions (Evans et al., 2022). For our study, we collected feedback from mathematics lecturers in response to a prompt on a common error made by undergraduate students. The task is derived from a digital module in integral calculus, designed to support undergraduates in finding areas enclosed by functions, as highlighted by Kontorovich and Locke (2023). Using a comparative judgment approach, two groups of judges – lecturers and students – will assess pairs of feedback for effectiveness. This will generate two scores for each item of feedback: one from lecturers and one from students. To address RQ1, we will compute the split-halves reliability measure for the group of lecturers, using the method described by Evans et al. (2022). This gives a number between 0 and 1, with values above 0.7 indicating a good level of agreement between the lecturers. To address RQ2, we will compute the correlation between the scores produced by the lecturers and the scores produced by the students. A high correlation (close to 1) indicates strong agreement on feedback effectiveness, while a low correlation (close to 0) suggests limited agreement.

If undergraduate students and mathematics lecturers share a common understanding of high- and low-quality mathematical feedback, this agreement motivates further investigation in two directions: evaluating alignment with established frameworks for high-quality feedback, and exploring potential differences in the underlying mechanisms and considerations influencing students and lecturers' choices. Conversely, any conflict in feedback perceptions between students and lecturers prompts an exploration of the factors contributing to the disparity. Addressing these factors is crucial for enhancing students' engagement with feedback.

References

Evans, C. (2013). Making Sense of Assessment Feedback in Higher Education. *Review of Educational Research*, *83*(1), 70–120. https://doi.org/10.3102/0034654312474350

Evans, T., Mejía-Ramos, J. P., & Inglis, M. (2022). Do mathematicians and undergraduates agree about explanation quality? *Educational Studies in Mathematics*, *111*(3), 445–467. https://doi.org/10.1007/s10649-022-10164-2

Kontorovich, I., & Locke, K. (2023). The Area Enclosed by a Function Is Not Always the Definite Integral: Relearning Through Collaborative Transitioning Within a Learning-Support Module. *Digital Experiences in Mathematics Education*, *9*(2), 255–282. https://doi.org/10.1007/s40751-022-00116-z

Lipnevich, A., & Panadero, E. (2021). A Review of Feedback Models and Theories: Descriptions, Definitions, and Conclusions. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.720195

Mulliner, E., & Tucker, M. (2017). Feedback on feedback practice: Perceptions of students and academics. *Assessment & Evaluation in Higher Education*, *42*(2), 266–288. https://doi.org/10.1080/02602938.2015.1103365

Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback: All that effort, but what is the effect? *Assessment & Evaluation in Higher Education*, *35*(3), 277–289. https://doi.org/10.1080/02602930903541007

Design of tasks for assessing Diophantine equations on the new Mexican School

Alejandra Fabiola Huitrado Mora[1] and Carolina de Haro Calderón[1]

[1]Teaching Updating Center in Zacatecas, Mexico; alejandrafabiola@camzac.edu.mx; carolina-calderon@camzac.edu.mx

Keywords: Diophantine equations, modeling, mathematical curriculum.

Linear Diophantine equations

In mathematics an equation is an algebraic statement in which it is shown that two amounts are equal using mathematical symbols like numbers, variable or unknown values joined by arithmetic symbols, and it is true for some values. This is an essential tool for modeling real life situations. Moreover, Amorim (2020) reminds us that one of the challenges in the initial training of mathematics teachers is to articulate the contents studied in the disciplines at the university with the themes of primary and secondary school. In this research we focus on the Diophantine equations as they are part of the new curriculum 2022, National Strategy for the Improvement of Normal Schools, in Mexico.

Given polynomial equations with integer coefficients and which solutions are also integer numbers are called Diophantine equations. For now, we focus on linear equations, more precisely with two variables, i.e., equations like the following: $ax + by = c$, with a, b and c integers numbers.

Curriculum 2022 National Strategy for the Improvement of Normal Schools

Recently, the basic education curriculum in Mexico has been updated with the aim of covering the needs of the development of individuals. For this, the new curriculum 2022 in its approach of the mathematics teaching and learning degree states that:

> Mathematics is understood as a complex social and cultural construction; On the one hand, it is a set of heterogeneous knowledge in permanent construction, dynamic and situated, and on the other, it is a scientific discipline with its own knowledge production procedures, which must be learned and taught among new generations. Its value lies in the fact that it allows the subject to situate himself, order and understand the world. (Secretaría de Educación Pública [SEP], 2022, p. 1)

Below I explain the treatment given to the design of the tasks for the evaluation of the topic "Diophantine equations".

Tasks-design for future teachers on the topic Diophantine equations.

Linear equations with one variable are perhaps the simplest tool for modeling and solving problems, but what happens when more than one unknown value appears? Moreover, what happens when we try to design activities to contextualize them, where there is also implicit information that the individual destined to pose and solve a problem must identify to use it in the correct modeling of the situation? A problem proposed for a group of 2nd semester of teachers in training to try to answer both questions was: On the "Pozolito" farm there are ducks and dogs. Mr. Genaro told his grandson that counting only the legs of the animals there are, there is a total of 54. Write an equation that describes the situation and finds at least one solution (if it is possible). Explain your answer.

The students took 15 minutes for solving the problem. Answering the first question, the most common error that occurred was that the implicit information (number of legs of both animals) was identified but the same variable was used, so the equation $2x + 4x = 54$ was proposed, and they quickly responded that reducing similar terms, i.e., $6x = 54$ the solution was $x = 9$. However, they do not differentiate between how many of them are ducks and how many of them are dogs. Later we will see, that 9 has no sense with the correct answers. Now for the second question, common errors were also identified. In this case, several of the students identified that the number of ducks was a variable x, and the number of dogs another variable y, but they did not use the implicit information (number of legs of both animals), so they proposed the equation $x + y = 54$, giving a long list of possible solutions, which was obviously incorrect because the implicit information was really fundamental for the statement and subsequent resolution. These difficulties were also reported by Edo (2013).

Then, what was the answer to the problem? As a first step, formulate the correct equation, that is, $2x + 4y = 54$, and then proceed to solve it. Clearly, the first thing to do is check that the equation does indeed have integer solutions, which are the ones that interest us and that they are clearly solutions to the contextualized problem, because even when we know that the equation has an infinite number of solutions in real numbers, we are not interested in saying that there are 3.5 dogs. It is easy to verify that this problem has integer solutions. Now, although in general we look for integer solutions for a Diophantine equation, and we have established that they are an infinite quantity, if they exist, for problems that model a situation of real life we must take in account which answers really are important to us, because in this case we are not interested in solutions like having -3 ducks either. Thus, the solutions reduce to $x = 1 + 2t, y = 13 - t, t \in Z, 1 \leq t \leq 12$.

The assessment of this problem, it was found that only a sixth of the group completely solved the task, and in total a quarter partially solved it, even intuitively, by trial and error. Due to all the factors observed, it is planned to replan the work with the contextualization of the contents, particularly with Diophantine equations, to improve future evaluations and the design of exercises paying attention in these and other factors that we can detect because they are an example of modeling simple real life.

Acknowledgment

To the Teaching Updating Center in Zacatecas for the supporting to this work.

References

Amorim, M. É., Pietropaolo, R. C., Galvão, M. E. E. L., & Da Fontoura G. S., A. (2020). A sequence of activities for teaching Diophantine Equations: Possibility to expand the knowledge base of future mathematics teachers. *Acta Scientiae, 22*(5), 207–225. https://doi.org/10.17648/acta.scientiae.6080

Edo, S. I., Ilma, R., & Hartono, Y. (2013). Investigating secondary school students' difficulties in Modeling Problems PISA-Model Level 5 and 6. *Journal on Mathematics Education, 4*(1). https://doi.org/10.22342/jme.4.1.561.41-58

Secretaría de Educación Pública [SEP] (2022, 03 March). *Anexo 12. Plan de estudio de la licenciatura en enseñanza y aprendizaje de las matemáticas.* 9FUnbRJp85-ANEXO_12_DEL_ACUERDO_16_08_22.pdf (sep.gob.mx)

Exploring instrumental orchestration practices in the context of formative assessment with technology

Min Chen[1], Rogier Bos[2], Michiel Doorman[3] and Paul Drijvers[4]

Freudenthal Institute, Utrecht University, The Netherlands; m.chen3@uu.nl;r.d.bos@uu.nl, m.doorman@uu.nl;p.drijvers@uu.nl

Keywords: Instrumental orchestration, formative assessment, digital technology.

Introduction

In guiding instrumental genesis teachers need to assess students' progress, and teachers' interventions involving digital tools can be described using the theory of instrumental orchestration (Trouche, 2004). We would like to study how assessment is part of these interventions of instrumental orchestration to better understand formative assessment in teaching with technology.

Background of the study

Digital tools have greatly increased the possibilities and potential for mathematical teaching and learning as well as for formative assessment (e.g., Baird et al., 2017). Much research on digital technology in mathematics education has been done with respect to learning with technology, and the significance of the instrumental view is central to current research on digital tools in mathematics teaching and learning (e.g., Drijvers et al., 2010; Trouche, 2004). Instrumentation theory stresses the importance of teachers carefully designing activities and selecting appropriate artefacts for students, to facilitate instrumental genesis in a meaningful and natural way. Regular whole-class teacher-centered teaching orchestration can be enriched with other formats, such as student-centered practices where students can actively explore mathematics activities through digital tools (Drijvers & Sinclair, 2023). These student-centered activities help to collect whole class information. Moreover, the integration of digital tools in the mathematics classroom requires careful consideration of how to orchestrate the lessons and the impact on assessment practices (Panero & Aldon, 2016). Thus, exploring teachers' orchestration preferences is crucial for aligning teaching practices with formative assessment using digital technology in mathematics classrooms.

Relation to the conference theme and dimension

According to Hollebrands & Okumuş (2018), effective instrumental orchestration involves selecting appropriate technological tools and guiding their use to foster meaningful learning experiences. In the context of formative assessment, technology can provide immediate feedback, facilitate personalized learning, and enable teachers to monitor student progress in real-time (Spector et al., 2016). This alignment of technology with pedagogical goals enhances the formative assessment process by making it more dynamic, responsive, and supportive of student learning trajectories (Black & Wiliam, 2009). Thus, the relationship between instrumental orchestration and formative assessment with technology could be synergistic, integrating digital tools to create a more effective and nuanced approach to evaluating and supporting student learning. This study adopts the instrumental orchestration theory to study the ways in which teachers' orchestrate formative assessment practices through digital technology. In other words, we want to study the role of formative assessment in how teachers develop orchestration strategies and implement them.

Setup of the study

The research question of this study is: How do teachers conduct formative assessment as part of their orchestration practices when using digital technology in mathematics classrooms?

We answer this question by studying classroom videos of technology-rich lessons. This study uses a convenience sampling method and participants are secondary school mathematics teachers from JiangSu Province in China. We choose videos of the 2022 "JiangSu Province ICT-integrated High-Teaching Quality Course Competition" for observation and analysis, because we expect that these videos could illustrate successful examples of technology-rich classroom practices. When analyzing the data, we focus on understanding how the orchestration types employed by teachers provide them with opportunities to gauge the progress of their students and to assess their work. The results examine teachers' orchestrations and classroom assessment practices with the aim to illustrate teachers' specific orchestration practices with a particular emphasis on assessing students' performance in a formative way. Hopefully, our findings can be used to demonstrate how teachers can create more adaptive and responsive learning environments that enhance student outcome by effectively orchestrating digital tools within formative assessment practices.

References

Black, P., & Wiliam, D. (2009).Developing the theory of formative assessment. *Educational Assessment, Evaluation & Accountability*,*21*, 5–31.https://doi.org/10.1007/s11092-008-9068-5

Drijvers, P., Doorman, M., Boon, P., Reed, H., & Gravemeijer, K. (2010). The teacher and the tool: Instrumental orchestrations in the technology-rich mathematics classroom. *Educational Studies in Mathematics*, *75*(2), 213–234. https://doi.org/10.1007/s10649-010-9254-5

Dalby, D., & Swan, M. (2019). Using digital technology to enhance formative assessment in mathematics classrooms. *British journal of educational technology, 50*(2), 832–845. https://doi.org/10.1111/bjet.12606

Drijvers, P., & Sinclair, N. (2023). The role of digital technologies in mathematics education: Purposes and perspectives. *ZDM,* 1–10. https://doi.org/10.1007/s11858-023-01535-x

Panero, M., & Aldon, G. (2016). How teachers evolve their formative assessment practices when digital tools are involved in the classroom. *Digital Experiences in Mathematics Education, 2*(1), 70–86. https://doi.org/10.1007/s40751-016-0012-x

Trouche, L. (2004). Managing complexity of human/machine interactions in computerized learning environments: Guiding students' command process through instrumental orchestrations. *International Journal of Computers for Mathematical Learning, 9*, 281–307. https://doi.org/10.1007/s10758-004-3468-5

Spector, J. M., Ifenthaler, D., Sampson, D., Lan (Joy) Yang, Mukama, E., Warusavitarana, A., ... & Gibson, D. C. (2016). Technology enhanced formative assessment for 21st century learning. *Journal of educational technology & society*, *19*(3), 58–71.

Hollebrands, K., & Okumuş, S. (2018). Secondary mathematics teachers' instrumental integration in technology-rich geometry classrooms. *The Journal of Mathematical Behavior*, *49*, 82–94. https://doi.org/10.1016/j.jmathb.2017.10.003

Self-assessment in long-term problem solving STEM contexts

Julia Schäfer[1] and Gero Stoffels[1]

[1]University of Cologne, Institute of Mathematics Education, Cologne, Germany;
julia.schaefer@uni-koeln.de; gero.stoffels@uni-koeln.de

Keywords: Self-assessment, problem solving, mathematics in STEM, authenticity.

MINTco@NRW a project fostering long-term problem solving

MINTco@NRW is a cooperation between the University of Siegen and the University of Cologne. This project is a follow up of the extracurricular project “Authentic-STEM” (Stoffels, 2024). Its aim is to integrate systematically long-term problem solving in authentic contexts into regular classrooms at secondary level in North Rhine-Westphalia. To make this possible, companies collaborate in the project and provide several authentic mathematics related unsolved problems that are solved by German and U.S. students in solver-teams. During a cycle the students work four months on the problems. At the end of each cycle, the solutions developed are presented to the companies. Through MINTco@NRW this activity will be transferred into regular classrooms. Therefore, there is a need for tasks and materials fitting the curriculum as well as testing adequately the participating students for grading. Both need to be connected to the core ideas of the project. These are, that through long-term and intensive engagement with authentic problems the students deepen their 21st century skills (Radmehr & Vos, 2020), enhance their mathematical and STEM competencies (e.g. problem solving, modeling, using theories and communicate) and explore their self-efficacy (Bandura, 1997).

Self-assessment for testing more than mathematical and STEM competencies

Various studies have already shown that there is a strong correlation between students' learning, performance, and formative assessments. Feedback and its quality also play a central role in the above-mentioned context. Feedback should point out the main errors and their probable causes and explain how to avoid those errors in the future. All three aspects can be addressed by students' self-assessment, since self-criticism and self-evaluation have a major influence on students’ own learning processes (Shepard, 2005). Also, self-evaluation plays a central role in many mathematical metacognitive processes, e.g. the verification phase in mathematical problem solving (Rott et al., 2021) or the validating phase in mathematical modeling (Blum & Borromeo-Ferri, 2009). The long-term approach of this project (Stoffels & Holten, 2022) as well as the complexity of the real problems, give many opportunities for the students to evaluate their own problem solving activity. Also, the setting with multiple stakeholders involved can add other perspectives to the students’ self-assessment. An example for this aspect is the evaluation of the students’ own views about the mathematicality of the problems. Similar questions are raised in the sociologically informed conceptualization of authenticity given by Vos (2018).

The main research questions are: how adequate self-assessment formats look like, which capture both more general skills and competencies as well as the quality of mathematical processes; and who, among the stakeholders, should take part in which way to strength the capability of students’ self-assessment? Therefore, during the project period, one focus of the self-assessment will lay on the students' zone of proximal development (Vygotsky 1978) also monitoring the project’s environment.

Methodological considerations and expected outcomes

The project consists of a half year pilot cycle in Spring 2024, which is still in an extracurricular format. During this cycle material (e.g. logbooks, Impulse formats) for supporting the students' problem solving are evaluated using structuring content analysis. Parallel to this identification formats for the self-assessment of students and tests for grading are identified and adapted to the core ideas of the project. In the following two years starting summer 2024, the program will be implemented in at least six regular classrooms in North-Rhine Westphalia, so two design research cycles will be carried out to improve the formats and evaluate their effectiveness. Over each cycle, portfolios are kept as "digital logbooks", which might be a candidate for examination formats.

The presented project aims to contribute to the development of adequate assessment formats for long-term cooperative problem solving as well as the evaluation of the adequacy of these formats regarding their ability to foster students' self-assessment and self-regulatory processes.

Acknowledgment

MINTco@NRW is funded by the Ministry of Education of North-Rhine Westphalia.

References

Bandura, A. (1997). *Self-Efficacy: the exercise of control*. W. H. Freeman and Company.

Blum, W., & Borromeo Ferri, R. (2009). Mathematical Modelling: Can It Be Taught And Learnt? *Journal of Mathematical Modelling and Application*, *1*(1), 45–58.

Radmehr, F., & Vos, P. (2020). Issues and challenges for 21st century assessment in mathematics education. *Science and mathematics education for 21st century citizens: Challenges and ways forwards*, 437-462 http://doi.org/10.13140/RG.2.2.18537.77927

Rott, B., Specht, B., & Knipping, C. (2021). A descriptive phase model of problem-solving processes. *ZDM–Mathematics Education*, *53*(4), 737-752. https://doi.org/10.1007/s11858-021-01244-3

Shepard, L. A. (2005). Linking formative assessment to scaffolding. *Educational leadership*, *63*(3), 66-70.

Stoffels, G., & Holten, K. (2022). MINT-Pro²Digi: Authentisches projektorientiertes mathematisches Problemlösen in außerunterrichtlichen digitalen Kontexten. *MINTUS – Beiträge Zur Mathematisch-Naturwissenschaftlichen Bildung*, 47-71. https://doi.org/10.1007/978-3-658-36764-0_3

Stoffels, G. (2024). *Authentic-STEM: Opening long-term domains of experience for fostering students' and mentors' selfefficacy through mathematics* (P. Drijvers, C. Csapodi, H. Palmér, K. Gosztonyi, & E. Kónya, Eds.). HAL Archives Ouvertes; Alfréd Rényi Institute of Mathematics. https://hal.science/hal-04420539

Vos, P. (2018). "How real people really need mathematics in the real world"—Authenticity in mathematics education. *Education Sciences*, *8*(4), 195. https://doi.org/10.3390/educsci8040195

Vygotsky, L. S., & Cole, M. (1978). Mind in society: Development of higher psychological processes. In *JSTOR*. Harvard university press. https://doi.org/10.2307/j.ctvjf9vz4

The role of elicitation in formative assessment

Kristoffer Arvidsson[1] and Torulf Palm[2]

[1]Umea University, Science and Mathematics Education, Sweden; kristoffer.arvidsson@umu.se

[2]Umea University, Science and Mathematics Education, Sweden; torulf.palm@umu.se

[1,2]Umea Mathematics Education Research Centre (UMERC)

Keywords: Formative assessment, elicitation, feedback, mathematics education.

Introduction and aim

Research has shown that formative assessment has a large potential for accomplishing positive effects on student learning (Black & Wiliam, 1998; Hattie, 2009; Andersson & Palm, 2017). One main function of formative assessment is to adapt teaching strategies to students' learning needs (SLN). Adaption of teaching to SLN requires that information about these needs be elicited and understood. However, studies examining feedback effectiveness most often do not focus on how the quality of the acquired information influences the impact of feedback on student achievement.

My poster will describe a study that is part of a larger project that seeks to contribute to the development of a theory of action for formative assessment. This includes identifying mechanisms by which formative assessment affects student learning. The study presented in this poster contributes to the project by identifying how a mathematics teacher elicits information about SLN when they seek help during task solving. The study aims to describe different ways of eliciting relevant information, how these ways influence the quality of information acquired and how these affect possibilities to adapt the feedback to meet the SLN.

This poster will describe the study and emphasise the importance of elicitation in the formative assessment practice. In doing so, describe the different ways of elicitation the teacher uses and how they may relate to the adjustment of feedback the teacher makes.

Method

This qualitative case study focuses on in-depth insights into the practice of one experienced middle school teacher. The data consists of 15 audio recordings of the teacher's mathematics lessons in a year 6 class over the span of one year. During this time the teacher was a part of a professional developmental program which focused on improving formative assessment practises in the classroom. During the recorded lessons, the teacher helped students individually when they sought help during task-solving. The data will be analysed in a thematic approach by identifying ways of elicitation, types of feedback and the relations between them.

Tentative findings

Preliminary findings suggest that the teacher's way of eliciting may influence the type and relevance of information gathered about SLN. The teacher's feedback seems to align more closely with SLN when the elicitation process yields sufficient information, i.e. to enable the adjustment of feedback specifically to those needs. In cases where elicitation does not provide sufficient information, but the teacher proceeds to give feedback, this feedback often misses addressing the actual learning need. Instead, it may focus on task completion, which could be a separate issue from the learning need.

Thus, it is important to explore how various elicitation ways might affect the outcome of elicitation and subsequently, how this outcome might shape the adaptation of feedback to students' learning needs.

Following are transcripts of two teacher-student interactions that depict two different ways of dealing with elicitation. The first transcript shows how the teacher elicits information about the learning need, continues to elicit information when she doesn't receive sufficient information and by the end of the conversation, the student's learning needs are resolved. The second transcript shows how the teacher elicits information but proceeds to give feedback before having received sufficient information about the learning needs. In the end, the learning needs of the student are not resolved, prompting the conversation to loop back, with the teacher needing to start again with the elicitation process.

Transcript 1

Student: I don't get it…
Teacher: What are you supposed to find out? What is the assignment? Can you tell me?
Student: I must figure out the circumference and area of this thing…
Teacher: OK, what help have you received [from the textbook] to be able to solve it?
Student: Ehh... all sides are 1 cm.
Teacher: Yes, OK, what about it do you find difficult?
Student: I don't know... I don't understand. Should I count all of them, like that?
Teacher: Yes?... What do you think... How do you calculate the circumference?
Student: But I'm going to... You add everything together.
Teacher: Yes, you already knew that! Great!

Transcript 2

Teacher: OK, what assignment are you working with?
Student: 128. I have done like this [shows previous work] and the thing I don't get is…
Teacher: [Interrupts] Wait, where are you? Explain the assignment.
Student: [Explains the assignment]
Teacher: What formula have you used?
Student: I don't know... or what do you mean?
Teacher: [Explains the formula for calculating area]
Student: I have used that formula already.
Teacher: Oh? I see, right... Then what was your problem?
[Conversation continues…]

A key distinction between these two ways of eliciting is the persistence in continuing until sufficient information is acquired. The poster session will offer a chance to explore additional examples.

References

Andersson, C., & Palm, T. (2017). The impact of formative assessment on student achievement: A study of the effects of changes to classroom practice after a comprehensive professional development programme. *Learning and Instruction*, *49*, 92–102. https://doi.org/10/f3t4k7

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10/fpnss4

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

FAME 1 Proceedings
Feedback & Assessment in Mathematics Education
5-7 June, 2024 - Utrecht, The Netherlands

14th ERME Topic Conference (ETC14)

ISBN: 978-90-393-7788-8
DOI: 10.5281/zenodo.14231455

Utrecht
University