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.

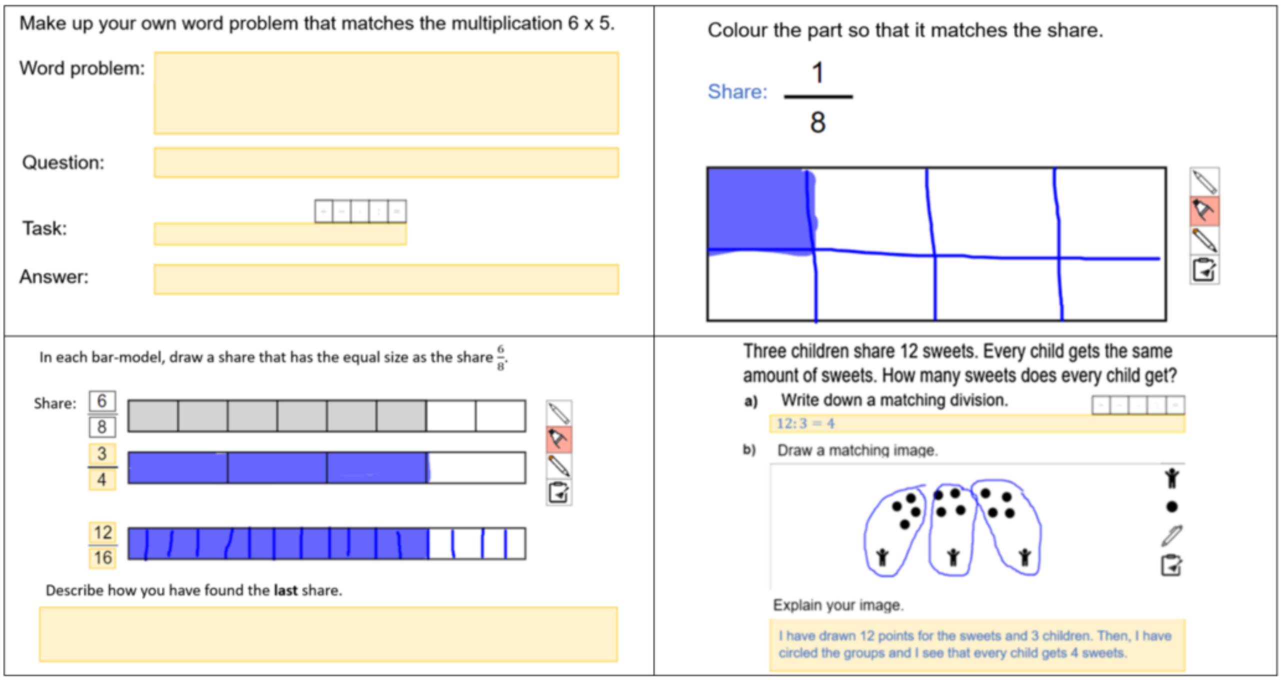
Typical challenges in coding open-ended items

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 $\frac{6}{8}$.

Share: $\frac{6}{8}$ $\frac{3}{4}$ $\frac{12}{16}$

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

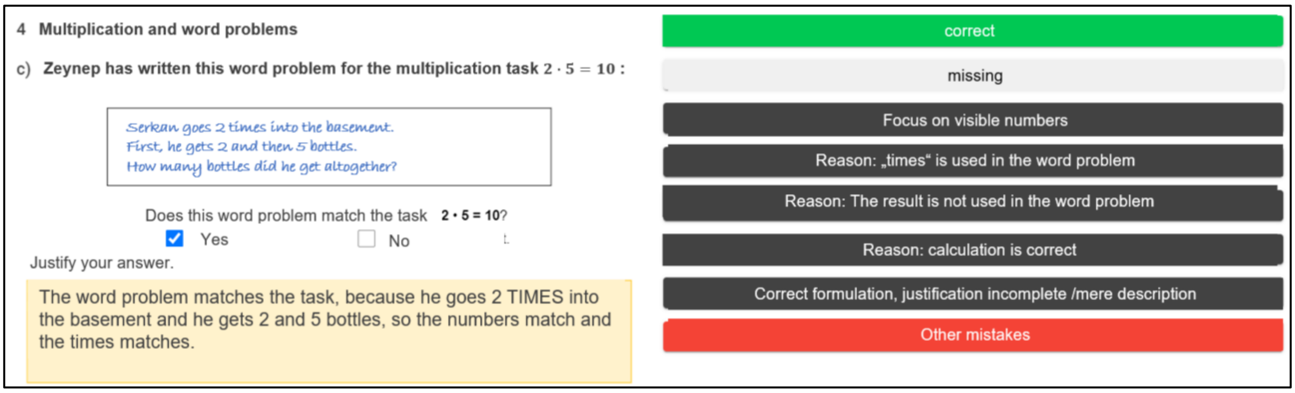
b) Draw a matching image.

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.

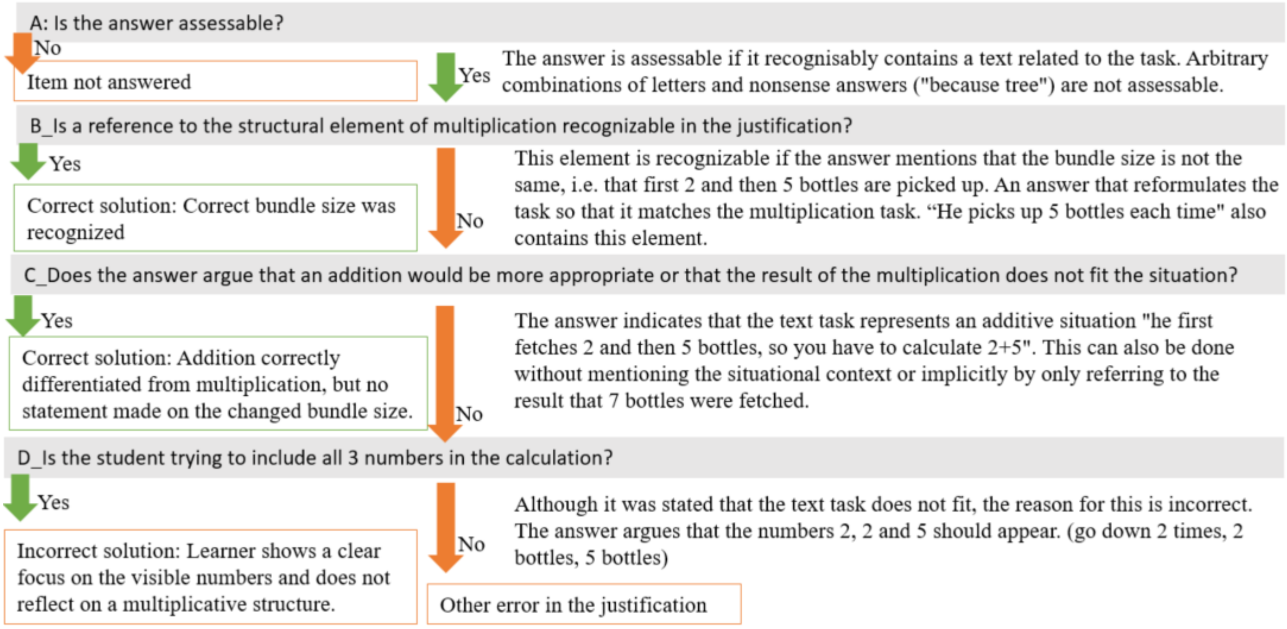
correct
missing
Focus on visible numbers
Reason: „times" is used in the word problem
Reason: The result is not used in the word problem
Reason: calculation is correct
Correct formulation, justification incomplete /mere description
Other mistakes

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 (κ > 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.

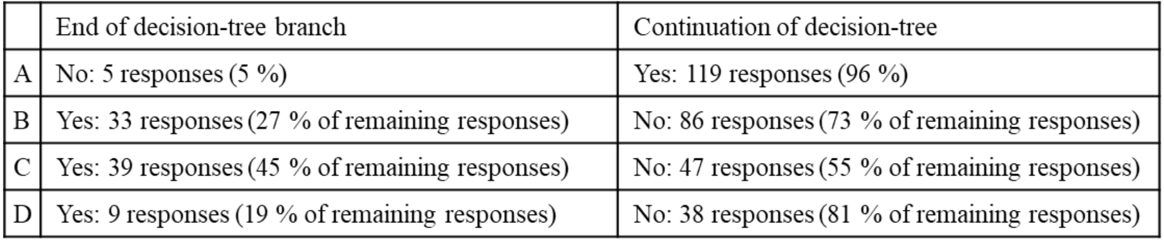
Empirical Study: Human and AI-coding with a decision tree

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

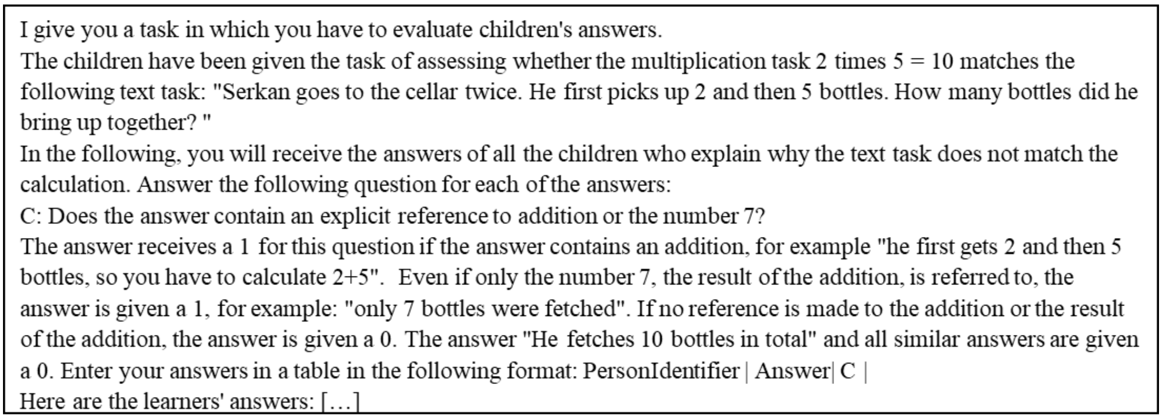
The expert-rating revealed that even though the hierarchical structure is only developed with respect to the assessed content, every group of responses is represented (Table 1) and only 38 responses (31 % of all responses) belonged to the “other error” category. 21 responses (55 % of this category) showed 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.