Handwritten math exams with multiple assessors: researching the added value of semi-automated assessment with atomic feedback

Filip Moons, Ellen Vandervieren

Handwritten math exams with multiple assessors: researching the added value of semi-automated assessment with atomic feedback

Filip Moons[1] and Ellen Vandervieren[1]

[1]University of Antwerp, Antwerp School of Education, Belgium; filip.moons@uantwerpen.be

Digital exams often fail in assessing all required mathematical skills. Therefore, it is advised that large-scale exams still feature some handwritten open answer questions. However, assessing those handwritten questions with multiple assessors is often a daunting task in terms of grading reliability and feedback. This paper presents a grading approach using semi-automated assessment with atomic feedback. Exam designers preset atomic feedback items with partial grades; next, assessors should just tick the items relevant to a student's answer, even allowing 'blind grading' where the underlying grades are not shown to the assessors. The approach might lead to a smoother and more reliable correction process in which feedback can be communicated to students and not solely grades. The experiment took place during a large-scale math exam organized by the Flemish Exam Commission, and this paper includes preliminary results of assessors' and students' impressions.

Keywords: Assessment, computer-assisted assessment, state examinations, feedback, inter-rater reliability.

Introduction

Regardless of all the practical advantages digital exams offer, Hoogland and Tout (2018) warn that digital questions focus on lower-order goals (e.g., procedural skills). They argue that handwritten questions are better suited to assess vital higher-order goals (e.g., problem-solving skills). Lemmo (2020) highlights substantial differences in students' thinking processes when the same question is asked digitally or paper-based. Bokhove & Drijvers (2010) point out that handwritten questions allow students to express themselves more freely. For all these reasons, Threlfall et al. (2007) advise deciding for each question individually whether the digital or handwritten mode is appropriate, leading to exams that are a mixture of both.

One major issue with handwritten questions is to find ways to assess them efficiently and reliably. Indeed, when the correction work is distributed among several assessors, guaranteeing grading reliability (Billington & Meadows, 2005) and consistent feedback (Baird et al., 2004) is challenging. Most exam designers try to ensure reliability by pre-developing a solution key with grading instructions for assessors (Ahmed & Polit, 2011).

General idea

In this paper, we introduce a novel approach to assess handwritten students' solutions with multiple assessors in a *semi-automated*[1] (SA) way: students solve these questions the classical way by writing on a sheet of paper. Next, these sheets are scanned, and assessors use the SA-system to correct the

[1] In the rest of this paper, we use the abbreviation **SA** to refer to our semi-automated assessment approach with reusable feedback.

solutions on a computer. Exam designers provide a solution key for each question consisting of different feedback items, written in an *atomic way* (see below), anticipating the most common mistakes. These feedback items can be linked to partial points for grading. When correcting a student's solution, the assessors have to select the appropriate feedback items, so the same feedback items are reused repeatedly. If a certain solution approach by a student is not covered in the available feedback items, an assessor can add a new feedback item. This new item is immediately available to all other assessors, leading to a *dynamic solution key* that expands as more and more corrections are made. When all assessors finish their job, the system produces individual reports for all students, including the grades and feedback.

In the following subsections, we discuss atomic feedback, the link with adaptive grading, and the idea of 'blind' grading.

Atomic feedback

Classic written feedback has traditionally consisted of long pieces of written text (Winstone et al., 2017). With its long sentences describing all of the errors in a student's work, classic written feedback is intrinsically not very reusable, as it is too explicitly targeted toward specific students. To overcome this difficulty and maximize the reusability of feedback, one of the key ideas underlying the proposed SA system is that it allows exam designers and assessors to write *atomic feedback* (see Figure 1). To write atomic feedback, one has to (1) identify the possible independent errors occurring and (2) write separate feedback items for each error, independent of each other. These atomic feedback items form a point-by-point list covering all items that might be relevant to a student's solution. The list can be hierarchical to *cluster* items that belong together (see Figure 1).

We have intensively studied atomic feedback in the first study of this PhD-project in the context of individual math teachers giving feedback to their students (Moons et al., 2020). The results of this study (Moons et al., in press) indicated that atomic feedback is significantly more reused than non-atomic feedback, so we found formal requirements to write feedback that can be reused. Teachers also tend to give more feedback when writing and reusing atomic feedback instead of saving time. However, since the atomic items are shared across multiple assessors in this second study, an additional criterium for atomicness is added: (3) a knowledgeable assessor must be able to determine unambiguously whether an item applies to a student's answer or not. As such, each item implicitly represents a yes/no question. Related atomic feedback items and intermediate steps in a solution key can share the same color to visually clear their connection (see Figure 1).

Adaptive grading

To obtain grades, exam designers can associate atomic feedback items with partial points to be added (green items in Figure 1) or subtracted. It is also possible to associate items with a threshold (e.g., 'if this feedback item is ticked, no points', red items in Figure 1).

The point-by-point list of atomic feedback items ultimately forms a series of implicit yes/no questions to determine the students' grade. Dependencies between items can be set, so items can be shown, disabled, or changed whenever a previous item is ticked, implying that assessors must follow the point-by-point list from top to bottom. This adaptive grading approach resembles a flow chart that

automatically determines the grade, but – by ticking the items that are relevant to a student's answer – might at the same lead to several other envisioned benefits: (1) a deep insight into how the grade was obtained for both the student (feedback) as well as the exam committee and (2) a straightforward way to do correction work with multiple assessors as personal interpretations are avoided as much as possible (inter-rater reliability).

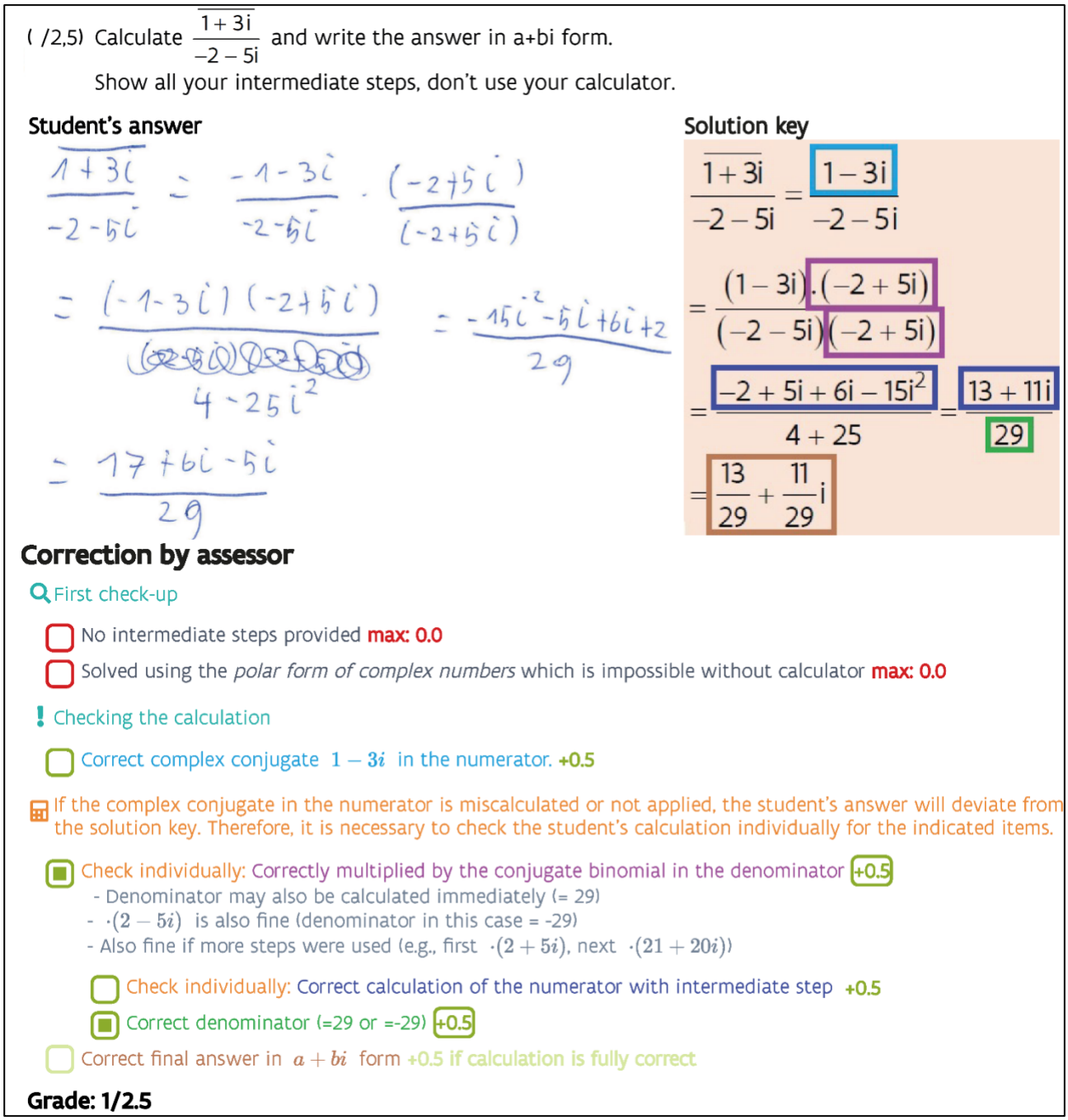

Figure 1: An example of adaptive SA grading with atomic feedback

In Figure 1, an example of the SA approach is given. The students' answer survives the 'First check-up' items; checking one of them would otherwise disable all of the following items. As the item 'Correct complex conjugate 1-3i' is unticked, the computer knows that a mistake happened; however,

assessors should continue their assessment of the answer by taking into account that the students' steps will now deviate from the solution key for some items; these items are indicated by 'Check individually.' All the orange content would have disappeared when the item 'Correct complex conjugate 1-3i' had been ticked. The item 'Correct final answer in a + bi form' only gets enabled when all previous green items are ticked. The two ticked items each add 0.5 points to the grade, leading to a total of 1 out of 2.5.

Blind grading

Imagine that all references to points/grades disappear in Figure 1. This leads to the experimental idea of '*blind grading*' where the assessor chooses the appropriate feedback items without seeing the associated scores. The system still calculates the grades, but these are invisible to the assessors. The envisioned advantage of this grading approach is that assessors only need to focus on the content of a student's answer; any emotional barrier to select a feedback item disappears, possibly leading to higher grading reliability. Indeed, Ahmed & Pollit (2011) already indicated that deviations from a traditional solution key often occur when assessors disagree with the obtained grade. A possible disadvantage is that assessors might be afraid of being too lenient or too harsh. They lose an important frame of reference since they cannot compare if the calculated grade matches their sense of fairness.

The opposite mode of *blind grading* will be called '*visible grading*' in the rest of the paper; this is the standard mode where assessors can see the associated points for every feedback item and the calculated total grade (see Figure 1). Note that blind grading should not be confused by anonymous grading (Hanna & Leigh, 2012); in anonymous grading, assessors do not see the student's names to avoid certain biases (e.g., gender, ethnicity,...).

Research questions

After introducing the general idea and the key concepts, we present the first two research questions associated with preliminary investigations on assessors' and students' impressions of SA grading.

(RQ1) How did assessors appreciate the SA system with blind/visible grading regarding perceived usefulness and ease of use?

(RQ2) How did the students perceive their personal atomic feedback with grades?

Methods & Materials

The experiment is being executed in association with the Examination Commission of the Flemish government. Flanders is the Dutch-speaking part of Belgium. Flanders is a region without any central exams (Bolondi et al., 2019): every secondary school decides autonomously on the assessment of students. Consequently, the Examination Commission does not organize national exams for all Flemish students but organizes large-scale exams for everyone who cannot, for whatever reason, graduate in the regular school system. This way, students who successfully pass all their exams at the Examination Commission can still obtain a secondary education diploma. Students participating in these exams prepare autonomously or use a private tutor/school. The Examination Commission only provides clear guidelines for students on the content of the exams, carries out all the exams, and

awards diplomas, but does not provide any teaching activities to students. We received ethical clearance from the ethical committee of the faculty of social sciences from the University of Antwerp.

Materials

Development of the 'group' SA-system

We developed an adaptation of the SA tool (described in Moons et al., in press) ready for handwritten assignments with a group of assessors. The tool is integrated as an advanced grading method in Moodle, an open-source e-learning platform. As Moodle is a framework offering many readily available components (such as a grade book, log in and uploading assignments,…), it guarantees rapid application development. The group assessment tool contains all the features explained in the introduction of this paper.

Mathematics exam

The mathematics exam for this experiment was developed by the exam designers of the Flemish Examination Commission in the way they always develop exams. Their solution key was turned to atomic feedback items for SA grading in close cooperation with us. The exam was one of the two math exams for the advanced mathematics track of Flemish secondary education and features complex numbers, matrices, space geometry, discrete mathematics, statistics, and probability. Interestingly, the exam is already a mixture of fully automated and handwritten questions: 46% of the exam grades are obtained with digital questions. Our experiment will only focus on the 54% part that consists of 10 paper-based questions with an open-answer format.

Survey based on the TAM model for assessors

We developed a short, validated survey based on the Technology Acceptance Model (Davis, 1989) to measure how assessors experienced the SA system's usefulness and ease of use. The survey distinguished between the SA system using visible grading and the SA system using blind grading.

Survey based on feedback perceptions for students

We constructed a questionnaire loosely based on Weaver (2006) to measure how students perceive the personal atomic feedback they received.

Participants

60 students participated in the math exam linked to this study. The grading work was distributed among the 3 exam designers (employees responsible for the math exams of the Flemish Examination Commission) and 7 external assessors. These external assessors are mathematics teachers across Flanders who do this as a side job.

Methods

The Examination Commission designed the exam in August 2021. Next, their correction key was transformed to atomic feedback for SA grading in close cooperation with us. In October, all assessors received training with the SA system using a demo exam. The students took the exam on the 29th of October, 2021. All their answers were scanned and made available in the SA system. Every assessor had one month to correct the exams with the SA system. All student's exams were distributed among

the assessors. Especially for this experiment, all assessors got 30 randomly selected exams on top that were corrected by all (assessors were not aware of this). Half of the assessors corrected the even question blind, the other half the odd questions. Correctors filled in the survey based on the TAM model when they finished their assessment work. At the end of November 2021, all exams were corrected, and a personalized survey was sent out to the students. In this survey, students got access to their solutions to the questions, together with the provided atomic feedback (see Figure 1). The survey probed students' understanding of their feedback, how they liked the atomic form of the feedback, and the usability of getting feedback along with grades. After completing the survey, the students were invited for an in-depth interview on the same topic as the survey.

In February 2022, all assessors will re-correct the 30 exams corrected by all, but this time in the traditional way of the Examination Commission. In this traditional way, they must just communicate a grade for each question based on a paper-based solution key. This re-assessment will give deep insights into the inter-rater reliability of (visible/blind) SA grading versus traditional grading.

Results & Discussion

The results of (RQ1) on the assessors' views measured using the TAM model are shown in Table 1. The scales are measured on a 7-point Likert scale.

Table 1: Results of the TAM model by the assessors for both visible as well as blind SA grading

Scales	**Visible SA grading** M±SD	**Blind SA grading** M±SD
1. Perceived Usefulness	5.7 ± 0.7	4.6 ± 1.5
2. Perceived Ease of Use	5.4 ± 1.0	4.5 ± 1.4
3. Anxiety	2.5 ± 1.1	3.6 ± 1.7
4. Attitude Towards Using	6.1 ± 0.8	4.4 ± 1.7
5. Behavioral Intention to Use	5.6 ± 1.2	4.4 ±1.7

Table 1 shows that assessors have a strong attitude towards using visible SA grading, meaning that they like working with the visible SA grading system. Assessors rated their anxiety for visible SA grading as low and gave a high rating to the perceived usefulness, perceived ease of use, and the behavioral intention to use for visible SA grading. Blind SA grading was less appreciated on all scales. All assessors (100%) indicated they preferred visible over blind grading. Reasons given include the lack of control (71.4%), an alienated feeling (57.1%), and fear of missing items to be ticked (42.8%).

For (RQ2) on the student's perceptions of their received personal feedback (see Figure 1 for an example), the corresponding survey items are listed in Table 2. Of the 60 students who took the exam, 36 students participated in this online survey (60%). Results are expressed on a 7-point Likert scale and indicate that they would greatly appreciate if the Examination Commission would adopt this approach. Students feel that they understand their atomic feedback, learn from it, and see the connection with the obtained grades. It is important to remember that these results are entirely based

on self-reporting, and other qualitative techniques (which are being carried out) are necessary to check if students indeed understand the given feedback.

Table 2: Overview of the students' survey items corresponding to their personal feedback

Students' survey item	M±SD
1. My feedback was too uninformative or brief to be helpful	3.6 ± 1.9
2. My feedback encouraged me to improve	4.7 ± 1.7
3. I will make even better exams based on my personal feedback	4.9 ± 1.6
4. This personal feedback helps me to reflect on what I have learned	5.0 ± 1.3
5. My feedback indicated clearly how my scores were calculated	5.5 ± 1.1
6. I understand most of my feedback	5.3 ± 1.4
7. It would be great if the Examination Commission always gave this type of feedback	6.3 ± 0.7
8. I feel demoralized or angry after reading my feedback	2.8 ± 1.8
9. The relationship between the feedback and the score is clear	5.2 ± 1.2

Conclusion

This paper introduced preliminary results of the second study of this PhD-project, investigating the possible added value of semi-automated assessment with atomic feedback when multiple assessors have to correct the same mathematics exam. The first results indicate that assessors rate visible SA grading highly but are less keen on using blind SA grading. On the other hand, students seem happy with the atomic feedback SA grading produces. Nevertheless, there are still many facets to this research study that have not been highlighted: the inter-rater reliability (comparison between blind SA, visible SA, and traditional grading), measurements for assessor reliability, the effect of the dynamic solution key,… are still uncultivated territory in the exciting universe of SA grading in mathematics education.

Acknowledgment

This research is funded by a PhD fellowship of FWO, the Research Foundation of Flanders, Belgium (1S95920N)

References

Ahmed, A. & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 259–278. https://doi.org/10.1080/0969594X.2010.546775

Bolondi G., Ferretti F., Santi G. (2019). National standardized tests database implemented as a research methodology in mathematics education. The case of algebraic powers. In 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* (pp. 4040-4056).

Utrecht, the Netherlands: Freudenthal Group & Freudenthal Institute, Utrecht University and ERME. https://hal.archives-ouvertes.fr/hal-02430515

Bokhove, C., & Drijvers, P. (2010). Digital tools for algebra education: Criteria and evaluation. *International Journal of Computers for Mathematical Learning*, *15*(1), 45–62. https://doi.org/10.1007/s10758-010-9162-x

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Baird J., Greatorex J. & Bell J. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 331–348. https//doi.org/10.1080/0969594042000304627

Hanna, R. & Leigh L. (2012). Discrimination in Grading. *American Economic Journal: Economic Policy*, *4*(4): 146–68. https://doi.org/10.1257/pol.4.4.146

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

Lemmo, A. (2020). A tool for comparing mathematics tasks from paper-based and digital environments. *International Journal of Science and Mathematics Education*, 1–21. https://doi.org/10.1007/s10763-020-10119-0

Billington, L., & Meadows, M. (2005). A review of the literature on marking reliability. *Report for the National Assessment Agency by AQA Centre for Education Research and Policy.*

Moons, F., Vandervieren, E. & Colpaert, J. (In press). Atomic, reusable feedback: a semi-automated solution for assessing handwritten (math) tasks? *Computers & Education Open*

Moons, F., & Vandervieren, E. (2020). Semi-automated assessment: The way to efficient feedback and reliable math grading on written solutions in the digital age?. In A. Donevska-Todorova, E. Faggiano, J. Trgalova, Z. Lavicza, R. Weinhandl (Eds.), *Proceedings of the Tenth ERME Topic Conference (ETC 10) on Mathematics Education in the Digital Age (MEDA),* 16-18 September 2020 in Linz, Austria, 393–400. https://hal.archives-ouvertes.fr/hal-02932218

Weaver M. (2006) Do students value feedback? Student perceptions of tutor's written respons. Assessment & Evaluation in Higher Education, 31(3), 379–394. https://doi.org/10.1080/02602930500353061

Threlfall, J., Pool, P., Homer, M., & Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics*, *66*(3), 335–348. https://doi.org/10.1007/s10649-006-9078-5

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist*, *52*(1), 17–37. https://doi.org/10.1080/00461520.2016.1207538