Semi-automated assessment: the way to efficient feedback and reliable math grading on written solutions in the digital age?

Filip Moons and Ellen Vandervieren

University of Antwerp, School of Education, Venusstraat 35, 2000 Antwerp, Belgium, filip.moons@uantwerp.be

Process-oriented feedback is powerful in math teaching yet highly labour-intensive. As a consequence, digital assessment with fully-automated feedback has received much attention. Despite this, digital assessment faces shortcomings concerning students' input: it only reads specifically-formatted answers, and learners solve higher-order questions more naturally when using paper-and-pencil. Therefore, we investigated the use of a semi-automated assessment (SA) method in which teachers write atomic feedback that can easily be reused for multiple students. SA is implemented in Moodle. During a lab study (n = 1800 corrections), we examined (1) whether SA saves time; (2) whether SA delivers more reliable scores compared to handwritten assessment; and (3) how teachers perceive SA. Mixed effect models were used for data analysis.

Keywords: digital assessment, feedback, semi-automated assessment, atomic feedback, reusable feedback.

INTRODUCTION

Process-oriented feedback is a crucial instrument in learning processes (Hattie & Timperley, 2007): it tells students which mathematical operations were appropriate (strengths), which were not (weaknesses) and how task solutions can be improved (strategies) (Rakoczy et al., 2013). To provide good process-oriented feedback, teachers need 'interpretative knowledge' (Mellone et al., 2020) of students' errors.

Fully-automated feedback (FA) versus paper-and-pencil assessment (PP)

Because it is time-consuming for teachers to provide such feedback, considerable research has been devoted to fully-automated assessment (**FA**) in mathematics education (Sangwin, 2013). FA provides extensive, immediate feedback for students, substantial time profits for teachers and often endless training possibilities as questions can be automatedly generated.

Less attention has been paid to overcoming the drawbacks of FA. First, preparation of FA questions is challenging as it is complex to create the accompanying correction schemes that give partial grades and adapted feedback. The central weakness is, however, the fact that not all mathematics questions can be easily automated; especially higher-order thinking questions are solved by students more naturally using paper-and-pencil (Threlfall et al., 2007). Furthermore, almost all digital test environments offer too little mathematical tools that allow students to express themselves mathematically, as they would with pen and paper. Answer possibilities are mostly limited to pre-defined response fields instead of free answering formats used in paper-and-pencil assessments (**PP**) (Kocher & Sangwin, 2016). Hoogland & Tout (2018) have shown

that, as a consequence, a lot of FA-questions focus on lower-order goals, such as procedural skills. Besides, for higher-order questions, no difference between the effect of immediate and delayed feedback is found yet (van der Kleij et al., 2015), tempering the need of FA for that kind of questions. All in all, FA does not easily allow to assess open-ended, challenging problems triggering higher-order thinking.

In mathematics, students' answers will contain systematic error patterns, meaning that different students often make analogous mistakes (Movshovitz-Hadar et al., 1987) and teachers keep on noticing the same mistakes again. As paper-based assessment still has an essential place in mathematics teaching (Threlfall et al., 2007), it is surprising that using these error patterns to speed up the assessment process remains largely unstudied in the literature. In the present research, we want to bridge the gap between FA and PP and develop a new, semi-automated assessment method (**SA)**.

Semi-automated assessment with atomic feedback (SA)

SA is a method in which students work out their solutions using paper-and-pencil, but the teacher assesses them digitally, making it different from PP-tests who are handwritten assessed by the teachers.

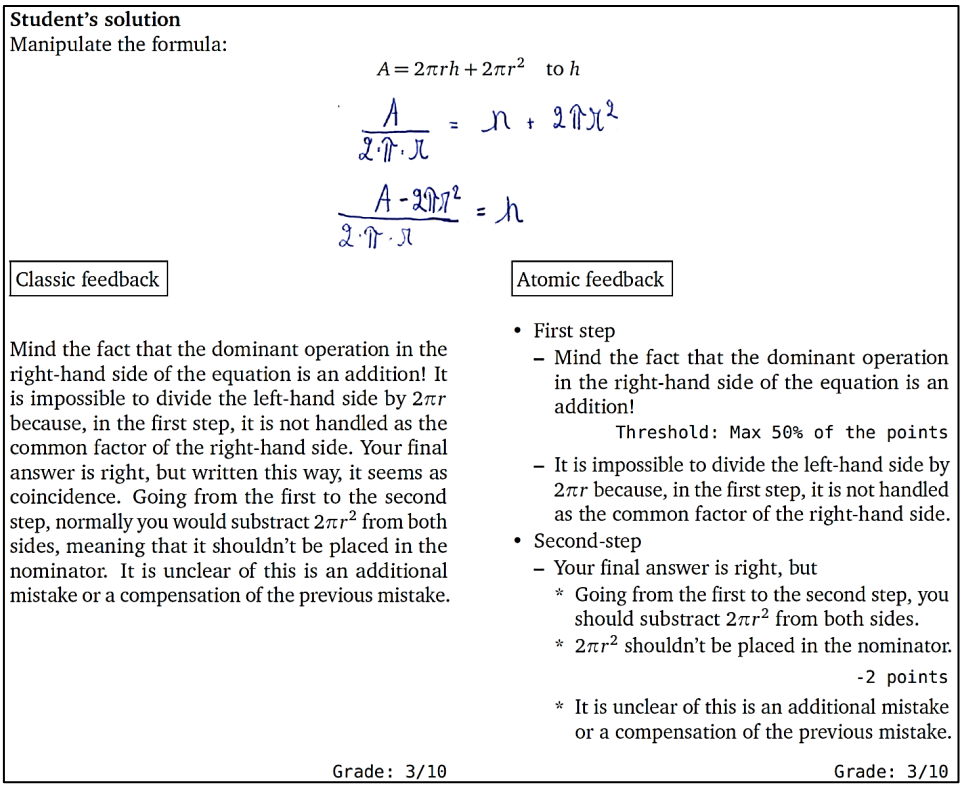

Figure 1: An example of classic versus atomic feedback

In SA, teachers have to write *atomic* feedback items. These feedback items are all saved, so they can easily be reused when another student makes the same mistake. The system suggests relevant items to reuse. The teacher can see the solutions of the students on-screen or asses directly from the students' sheets. It is possible to assess

test-by-test or question-by-question. SA generates a student report that can be printed or viewed online in an e-learning system.

To achieve high levels of reusability in SA, teachers must give *atomic* feedback: instead of writing long pieces describing lots of different mistakes at once, they must (1) identify the independent error occurring, and (2) write small feedback items for each error, independent of each other (see Figure 1). As such, SA can create point-by-point feedback only covering those items that are relevant to a student's solution. In addition, *clustering* of feedback is allowed, meaning sub-items can be added to feedback items. Clustering ensures that feedback can be written as atomic as possible and avoids teachers' need to write too specific items, which would compromise the reusability. It also allows related feedback to be orderly shown to students. Besides, the feedback cluster will be a decisive factor for the algorithm to decide which feedback will be suggested for reuse to the teacher. However, to support maximal flexibility, a feedback item can be part of different clusters.

If a question must be graded, the teacher can associate feedback items with partial scores to be subtracted. It is also possible to associate items with a threshold (e.g. 'if this feedback item is given to a student, a student can get at most 50% of the points'). The teacher can always still manually change the associated score of an item.

Solutions to assess handwritten students' tests digitally are available, like *Gradescope* (Singh, A. et. al., 2017), but the integration of reusable atomic feedback items has never been studied before.

Envisioned benefits of semi-automated assessment

SA might be a promising go-between for FA and PP, throwing off the current limitations of FA. First, SA gives rise to potentially significant time savings: solutions are assessed reusing already given feedback as much as possible. This might enable a faster feedback and grading process than PP, especially when a question has already been assessed many times, filling the database with lots of reusable feedback. Second, SA allows students to write down any mathematical expression, using the structure they prefer for their reasoning, and hence fully expressing themselves mathematically. Third, SA does not limit the use of open-ended, challenging higher-order thinking questions as there are no pre-defined response fields; the assessment work is in the hands of the teacher. Fourth, a teacher only gives feedback when a mistake occurs (no need for a crystal ball), omitting the need to develop complex correction schemes beforehand as is the case for FA (Sangwin, 2013). The loss of immediate feedback is a drawback, but remember that no significant difference in effect is found yet between delayed and immediate feedback for higher-order thinking questions (van der Kleij et al., 2015). In many cases, SA assessment might thus be a valid assessment method, combining the strengths of PP- and FA-assessment (see Table 1).

This study aims to verify these envisioned benefits experimentally. As we are currently collecting data, we have organised the rest of this proposal as follows: we introduce

the guiding research questions, examine the followed methodology and conclude by looking at possibilities for further research.

Feedback & Assessment		
	Computer-assisted	
Paper-and-pencil based (PP)	**Semi-automated (SA)**	**Fully automated (FA)**
- delayed feedback	- delayed feedback	+ immediate feedback
+ natural mathematical expressions	+ natural mathematical expressions	- too little mathematical tools
+ no pre-defined response fields	+ no pre-defined response fields	- pre-defined response fields
+ questions are easy to develop	+ questions are easy to develop	- need for an automated correction scheme and anticipation on mistakes
+ high-order thinking questions possible	+ high-order thinking questions possible	- high-order thinking questions difficult
- time consuming	+ time profits	+ time profits

Table 1: Advantages and disadvantages of different types of assessment

RESEARCH QUESTIONS AND HYPOTHESES

(RQ1) Does SA-feedback lead to significant time savings compared to PP-feedback and can we predict them?

Hypotheses: *As SA-feedback and SA-grades are reusable, we hypothesise that SA will be faster than PP. It might be possible that SA only becomes faster when a certain threshold (cf. number of tests assessed) is reached, as the database first must be filled with reusable feedback. Before that threshold, we hypothesise that there will be no significant time difference between SA and PP. To predict possible time savings, we seek for reusability measurements of the used feedback items, e.g. the ratio of already used feedback items to all feedback items used to correct a student's solution. When a solution is assessed with exclusively new feedback items, this ratio will be 0. If assessing a solution requires 5 different feedback items, of which 4 have already been used before, the reusability factor equals 0.8.*

(RQ2) Is teachers' SA grading more reliable compared to PP-grading?

Hypotheses: *Reliability is the degree to which an assessment produces stable and consistent results (Feldt, 2004). We focus on intra-rater reliability: how consistent is a teacher's grading (for a particular set of tests) over time? Extensive research has shown that teachers' PP-assessments are biased in numerous ways as teachers tend to forget how they handled the same mistakes before (Parkes, 2012). Because SA remembers already given feedback and associated grades, we hypothesise that the SA-grading stability will be better. With respect to inter-rater reliability, we expect no significant differences between SA and PP-grading, because in the current experiment teachers only use their own atomic feedback items. However, a follow-up study with a group of assessors contributing to and sharing the same database of atomic feedback items, is planned.*

(RQ3) A) How do teachers perceive SA? B) What about the feedback quality in the different conditions?

Hypotheses: *We hypothesise that teachers will appreciate the way SA integrates in their classroom practice as opposed to FA, which can sometimes feel a bit alienating. However, learning to write atomic feedback and using the SA-tool might be difficult. We will also compare the quality of given PP- and SA-feedback. As teachers are constantly remembered of already given feedback under SA, this could increase their interpretative knowledge (Mellone et. al., 2020).*

METHODS & MATERIALS

Materials

Development of MathSA

We developed an SA-tool called 'MathSA', and integrated it as an advanced grading method in the open-source e-learning platform Moodle. The Moodle-framework contains a lot of features (e.g. a grade book, uploading assignments,…) and is the most popular e-learning platform.

Test on linear equations

In close cooperation with a math teacher, we developed a test on linear equations, consisting of three equally weighted questions: (1) solve an equation (easy/procedural), (2) manipulate a formula (complex/procedural, see Fig. 1) and (3) a modelling question consisting of a word problem (complex/problem-solving). The three questions combined form a representative, standard test on linear equations.

Survey based on the TAM-model

We will develop a short, validated survey based on the Technology Acceptance Model (Davis, 1989) in order to measure how teachers perceive SA.

Participants

60 students of Grade 9 in one secondary school in Flanders (Belgium) solved the test on linear equations. We gathered informed consents from all students.

30 Belgian secondary math teachers with at least 3 years of working experience will participate voluntarily in a lab study. They were contacted through announcements in math teaching magazines and subscribed via www.mathsa.uantwerpen.be. We aim for diversity among participating teachers in gender, experience and school type. We plan to organise a focus group on MathSA and atomic feedback with 8 of the participating teachers. They will be selected based on their answers in the survey (cf. diversity in terms of gender, experience and views on technology). We received ethical clearance.

Design

All students conducted the test on linear equations in an authentic context: the students had been studying linear equations during classes, they were accustomed to the test lay-out, and afterwards, the grades were incorporated in the students' grade reports.

During the lab study, each teacher will assess all 60 solved tests. For each teacher individually, a random selection of 30 tests will be assessed under the SA-condition. The remaining 30 tests will be assessed under the PP-condition. This yields 1800 test corrections in total and indicates a within-subject design. With respect to RQ2, we will ensure every test is marked the same number of times under each condition. To avoid bias coming from a growing familiarity with the test, exactly half of the teachers will start with assessing PP-tests; the other half will first handle their SA-tests.

During the whole study, participating teachers will not be informed about the research questions, to prevent bias in their grading style. To control for bias due to inexperience with MathSA (cf. SA-condition), we provide sufficient training opportunities during the lab study, before we start with the actual experiment.

In RQ1, the dependent variable is the time a teacher needs to assess a single question. The independent variable is the assessment condition (PP/SA). As the assessment time also depends on: the teacher (categorical), the quality of the student's answer (measured by the test score), and the familiarity the teacher has with the test items (number of 1 to 30, indicating how many tests the teacher has already corrected under the same condition), these are all included as moderating variables.

For RQ2 (reliability of SA-grading), at least one month after the lab study, teachers will be asked to grade the same tests again under the same condition. This period in between guarantees that teachers will largely have forgotten how they handled particular tests. We will calculate the differences in scores between both measurements (score lab study – score month after) and use this as the dependent variable. The assessment condition (PP/SA) will be used as the independent variable. We will include the teacher (categorical) and the quality of the student's answer (measured by the average of the test scores given by all the teachers during the lab study) again as moderating variables.

To answer RQ3a, we will survey the participating teachers and conduct a focus group. To compare the given feedback under both conditions and whether they show different levels of interpretative knowledge (RQ3b), text mining will be used.

Procedure

In February 2020, 64 students (9th grade) solved the test on linear equations during their regular math class. It was conducted like every other test by their teacher and solved with paper-and-pencil. The researchers were not present during this test taking. Students were asked afterwards if the test could also be used for the research. All students agreed. We randomly deleted 4 tests to have exactly 60 tests.

We have been conducting the lab study on different moments in the months of July and August 2020. Due to the Covid-19 crisis, it wasn't feasible to gather all the 30 teachers at the same time in one place. During the lab study a presentation about useful process-oriented feedback in mathematics was given. Second, teachers were trained to get familiar with MathSA and the formulation of atomic feedback. Before the start of the actual experiment, all the teachers got half an hour to get familiar with the MathSA-tool. We offered some students' answers on entirely different math topics than the topic of the test of the experiment as training possibilities. During this training, teachers got tips to make good atomic feedback and had the opportunity to ask questions. Third, half of the teachers started assessing under the PP-condition: they wrote detailed feedback on each test and graded it. They were allowed to develop a personal correction scheme in advance. Every time they started to asses a test, they had to push the space button on the computer in front of them, so that the time needed to correct the test could be tracked. The other half of the teachers started assessing under the SA-condition, providing atomic feedback and scores with MathSA. The tool automatically keeps track of the time used for each test. In both conditions, participants were never allowed to return to an already corrected test. Fourth, after the break, the groups swapped conditions (SA/PP) and assessed the other, remaining tests. Finally, they were asked to fill in the survey.

After all the lab experiments on different dates are executed, we will organise the focus group with 8 participants online at the end of August.

In September 2020, a month after the lab study, the participants will receive the ungraded copies of their 30 PP-tests by post. They will be asked to re-grade them and will be invited to re-grade the remaining 30 SA-tests online. Their previous SA-corrections will have disappeared, but their feedback items of the lab study will have been saved. They will have one month to re-score the 60 tests under the same condition as during the lab study. Their PP-grades will be sent back to us through an online form. At the start of the experiment, all participants have been informed about this additional individual work (about 3 hours), but they do not know that they must re-grade the same tests.

Data analysis

We will construct mixed models (i.e. models containing both fixed effects and random effects) to examine the time differences (RQ1) and the consistency differences (RQ2) between SA and PP. In both models, the fixed effect is the condition (PP/SA). The random effects are the moderating variables mentioned in the design. The survey data will be analysed and cross-tabulated with teachers' characteristics (e.g. age, experience, technology acceptance score). We will also link the survey with the qualitative data from the focus group.

FURTHER RESEARCH

This paper describes the first study of this doctoral research. The goal of this first study is to explore SA as a new assessment method and get an indication on how SA behaves when teachers use it. The next step of the project is to focus on the students' point of view and conduct quasi-experimental studies to measure students' learning effects. We also plan an integration of SA-assessment with Bayesian networks for elaborate student tracking.

REFERENCES

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–340.

Feldt, L. S. (2004). Estimating the Reliability of a Test Battery Composite or a Test Score Based on Weighted Item Scoring. *Measurement and Evaluation in Counseling and Development*, *37*(3), 184–191.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*, 81–112.

Hoogland, K., & Tout, D. (2018). Computer-based assessment of mathematics into the twenty-first century: pressures and tensions. *ZDM*, *50*(4), 675–686.

Kocher, N., & Sangwin, C. (2016). Automation of mathematics examinations. *Computers & Education*, *94*, 215–227.

Mellone, M., Ribeiro, M., Jakobsen, A., Carotenuto, G., Romano,P. & Pacelli, T. (2020): Mathematics teachers' interpretative knowledge of students' errors and non-standard reasoning. *Research in Mathematics Education*, 1479-4802

Movshovitz-Hadar, N., Zaslavsky, O., & Inbar, S. (1987). An Empirical Classification Model for Errors in High School Mathematics. *Journal for Research in Mathematics Education*, *18*(1), 3–14.

Parkes, J. (2012). Reliability in Classroom Assessment. In J. H. McMillan (Ed.), SAGE Handbook of Research on Classroom Assessment. SZAGE, 107–124.

Rakoczy, K., Harks, B., Klieme, E., Blum, W., & Hochweber, J. (2013). Written feedback in mathematics: Mediated by students' perception, moderated by goal orientation. *Learning and Instruction*, *27*, 63–73.

Sangwin, C. (2013). Computer-aided Assessment of Mathematics. Oxford University.

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.

Singh, A., Karayev, S., Gutowski, K., & Abbeel, P. (2017). Gradescope: a Fast, Flexible, and Fair System for Scalable Assessment of Handwritten Work. *Proceedings L@S '17.*

van der Kleij, F. M., 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.