# Blind versus visible checkbox grading: Does not seeing the grades when assessing mathematics enhance inter-rater reliability?

Filip Moons[1,2] and Ellen Vandervieren[2]

[1]Utrecht University, Freudenthal Institute, the Netherlands; f.moons@uu.nl

[2]University of Antwerp, Antwerp School of Education, Belgium

*Assessing handwritten mathematics questions with multiple assessors is a daunting task regarding grading reliability. One of the primary sources of assessor variability arises when the grade resulting from the correction scheme does not align with their holistic appreciation of the student's answer. In this study, we developed a 'checkbox grading' approach that could possibly circumvent this variability source. In our approach, exam designers preset atomic feedback items with partial grades; next, assessors should tick the items relevant to a student's answer, allowing 'blind grading' when the underlying grades are not shown to the assessors. The study was executed in cooperation with the Flemish Exam Commission with 10 assessors and 30 students during a mathematics exam. This paper answers the question 'Does blind checkbox grading enhance inter-rater reliability?' In order to do so, we compared 'blind grading' with 'visible grading' and concluded that blind grading enhances the inter-rater reliability when the grading scheme is stringent.*

*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 (2021) 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, it is best to decide for each question individually whether the digital or handwritten mode is appropriate, leading to exams that are a mixture of both (Threlfall et al., 2007).

One major issue with handwritten questions is finding 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 & Pollit, 2011).

However, pre-developed solution keys are not perfect. From the literature on rubrics (Doğan & Uluman, 2017) we know that one source of assessor variability emerges when assessors' *holistic* grade differs from the *calculated* grade. The holistic grade is the grade they intuitively want to give when scoring a student's product (e.g., a math exam question) whereas the calculated grade is obtained when following the scoring guidelines from the rubric criteria they selected for the product.

When the calculated grade does not align with their holistic appreciation of the work, assessors often start changing the selection of criteria, which compromises the instrument's reliability (Dawson, 2017).

In this paper, we introduce 'checkbox grading': an assessment method for handwritten mathematics questions that can possibly overcome this source of assessor variability. In the following subsections, we discuss this method, the idea of 'blind' grading and the research question.

**Checkbox grading**

'Checkbox grading' is a semi-automated way to assess handwritten students' solutions with multiple assessors (e.g., high-stakes mathematics exams): students solve questions the classical way by writing on a sheet of paper. Next, these sheets are scanned, and assessors use an online system to assess the solutions on a computer. Exam designers provide a solution key for each question consisting of different feedback items written in an *atomic way* (Moons et al., 2022), anticipating the most common mistakes. These feedback items can be linked to partial points that will be added (green items in Figure 1) or subtracted (red items in Figure 1), or linked to a threshold for grading (e.g., 'if this feedback is checked, no points). When correcting a student's solution, assessors must select the appropriate feedback items (selecting the 'checkboxes'), so the same feedback items are reused repeatedly. When all assessors have finished their job, the system produces individual reports for all students, including the grades and feedback (the selected checkboxes are then filled with the appropriate color), like the one in Figure 1.

The point-by-point list of atomic feedback items ultimately forms a series of implicit yes/no questions to determine the student's grade. Dependencies between the checkboxes can be set, so that items can be shown, disabled, or changed whenever a previous checkbox 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 – ticking the items that are relevant to a student's answer – might at the same time 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).

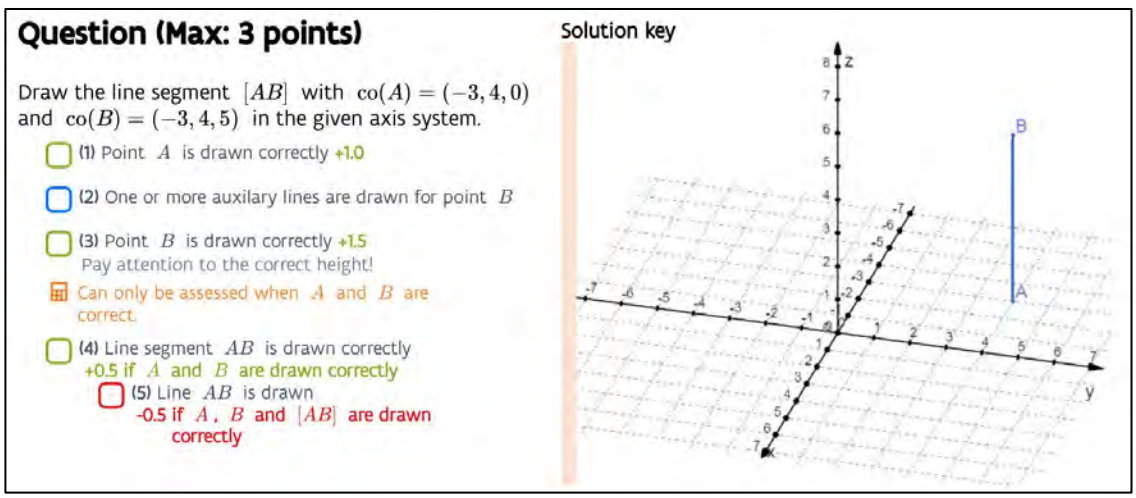


**Figure 1: An example of checkbox grading**

In Figure 1, an example of the 'checkbox grading'-approach is given. With this drawing question, a student can gain a maximum score of 3 points. If point *A* (1st box) is drawn correctly, the student gains 1 point; the correct drawing of point *B* is worth 1.5 points (3rd box). The second box does not change the score but shows assessors that the presence of auxiliary lines is perfectly fine. The last two feedback items, checkboxes 4 and 5, can only be selected if the first two have been selected. As the drawing of line *AB* implies the drawing of the line segment *[AB]*, the 5th box can only be selected if the 4th has been selected. The 5th box punishes students with -0.5 points if instead of line segment [AB], the line through A and B is drawn.

**Blind grading**

Imagine that all references to scores disappear in Figure 1. This leads to the experimental idea of 'blind grading' where assessors choose 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 selecting a feedback item disappears, possibly leading to higher grading reliability (Ahmed & Pollit, 2011). Moreover, it removes the conflict that can arise between the holistic and calculated grades of the assessor (Doğan & Uluman, 2017; Dawson, 2017). On the other side, a possible disadvantage is that assessors might fear being too lenient or harsh. By making the grades invisible, 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 with anonymous grading (Hanna & Leigh, 2012); in anonymous grading, assessors do not see the students' names to avoid certain biases (e.g., gender, ethnicity).

## Research question & framework

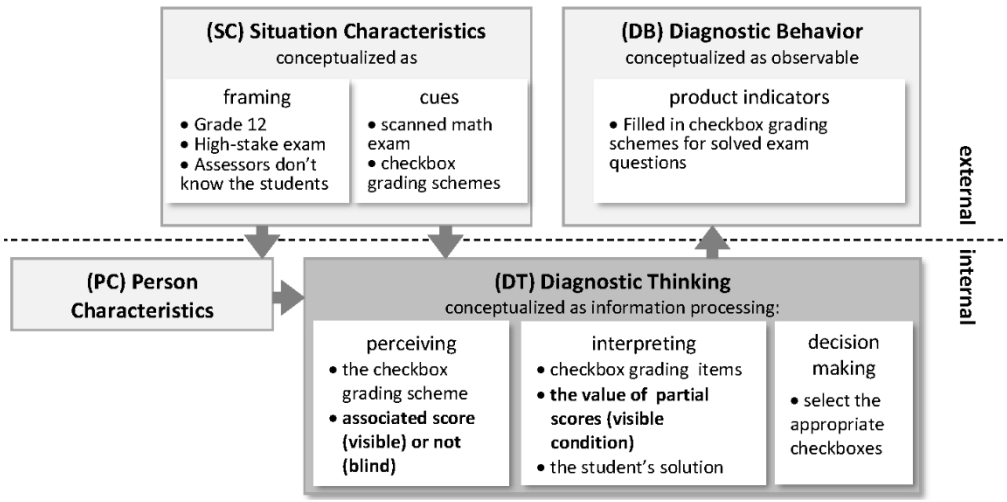


**Figure 2: Cognitive model for diagnostic judgments applied to our study (DiaCoM, Loibl et al., 2020).**

The research question central to this paper is: 'Does blind checkbox grading enhance inter-rater reliability compared to visible checkbox grading?' To frame our research we used the DiaCoM

framework (Loibl et al., 2020). The DiaCom framework describes teachers' judgments when assessing students' performance. It separates the assessment process into four components: the *Situation Characteristics* (e.g., grade level or subjects), *Person Characteristics* (e.g., teachers' states and traits), *Diagnostic Thinking* (i.e., teachers' decision making based on their perceiving and interpreting of information), and teachers' *Diagnostic Behavior* (e.g., teacher judgments). More specifically, we changed their perceiving and interpretation of their *diagnostic thinking* by switching between visible and blind grading, and looked how it changed their *diagnostic behaviour*. The DiaCom framework for this study is displayed in Figure 2.

## Methods & materials

The study was executed in association with the Exam Commission of the Flemish government. Flanders, the Dutch-speaking part of Belgium, is a region without any central exams: every secondary school decides autonomously on the assessment of students. Consequently, the Exam Commission does not organise national exams for all Flemish students. However, it organises large-scale, standardised exams for everyone who cannot, for whatever reason, graduate in the regular school system. This way, students who pass all their exams at the Exam Commission can still obtain a secondary education diploma.

The mathematics exam for this experiment was developed autonomously (without any influence from the researchers) by the exam designers of the Flemish Exam Commission in the way they always develop their exams. The exam is the second of two exams to pass the advanced mathematics track in the 12th grade. Their solution key was turned into checkbox grading in close cooperation with us. The handwritten exam was one of the two math exams for the advanced mathematics track of Flemish secondary education and featured complex numbers, matrices, space geometry, statistics, and probability. An overview of the question and their scores can be found in Table 1. The questions vary considerably in points that could be gained, based on the importance of the topic in the curriculum and the complexity of the question; 0.5 points was the smallest possible partial score.

| Question | Topic | Learning goal | Max. Score | M ± SD |
|---|---|---|---|---|
| Q1 | Complex numbers | Calculations with complex numbers in $a + bi$-form. | 2.5 | 1.75 ± 0.88 |
| Q2 | Complex numbers | Calculations with complex numbers in polar form | 2.5 | 0.67 ± 0.61 |
| Q3 | Matrices | Modelling with matrices | 3.5 | 1.95 ± 0.96 |
| Q4 | Matrices | Coefficient matrices of linear equations | 3.5 | 1.18 ± 0.96 |
| Q5 | Space geometry | Parameter equations of a plane | 1.5 | 0.18 ± 0.42 |
| Q6 | Space geometry | Cartesian equation of a line | 1 | 0.04 ± 0.20 |
| Q7 | Space geometry | Drawing a segment line in the $x, y, z$-assenstelsel | 2.5 | 1.16 ± 0.80 |
| Q8 | Space geometry | Determining the distance between a point and a line | 4.5 | 0.57 ± 1.35 |
| Q9 | Space geometry | Parallel lines in space geometry | 2.5 | 0.76 ± 0.94 |
| Q10 | Statistics & Probability | Modelling a probability experiment | 4 | 0.39 ± 1.13 |
| **WHOLE EXAM** | Algebra – Geometry – Statistics & Probability | | **28** | **8.65 ± 4.93** |

**Table 1: Content of the mathematics exam, including the scores**

The exam, including the checkbox grading schemes, can be found in Moons (2021).

Sixty students took the math exam linked to this study. The grading work was distributed among the three exam designers and seven external assessors. These external assessors are mathematics teachers across Flanders who do this as a side job. From these sixty students, we selected all exams containing a maximum of two questions left empty. From this selection, we randomly drew 30 exams that would be assessed by all 10 assessors (exam designers + external assessors). Half of the assessors corrected the even questions blind, and the other half the odd questions.

## Results

To measure the inter-rater reliability, we calculated a chance-corrected kappa (Moons & Vandervieren, 2023) for every question, for the whole exam, and separate $\kappa$ values for each condition. The kappa-statistic is a generalisation of the Fleiss' kappa. It varies between -1 and 1, with 1 indicating perfect agreement, 0 indicating no agreement better than chance, and a value below zero indicating the agreement was less than one would expect by chance. The kappa-statistic measures the inter-rater reliability and considers the feedback that is selected, the partial scores and the dependencies among the checkboxes.

In order to answer the research question, we used bootstrapping with 10,000 bootstrap samples for each question (and the whole exam) to test if the differences in $\kappa$ values of both conditions are statistically significant ($H_0: \hat{\kappa}_{blind} - \hat{\kappa}_{visible} = 0$). As each condition consisted of a different group of assessors (linking to the even/odd treatment), we used an unpaired bootstrap hypothesis test. An overview of this analysis for each question and the exam as a whole can be found in Table 2. Along with the significance test, we also used 10,000 bootstrap samples for every $\kappa$-value reported in Table 2 to determine the bootstrap 95% confidence intervals. We could not use a classic statistical test as a general expression of the theoretical sampling distribution of the $\kappa$-statistic is still lacking in the literature.

| **Question** | **Overall** | | **Blind grading** | | **Visible grading** | | ***p*-value** |
|---|---|---|---|---|---|---|---|
| | $\kappa$ | 95% CI | $\kappa$ | 95% CI | $\kappa$ | 95% CI | |
| Q1 | 0.803 | (0.72 to 0.90) | 0.833 | (0.75 to 0.94) | 0.767 | (0.66 to 0.89) | .185 |
| Q2 | 0.641 | (0.54 to 0.77) | 0.812 | (0.72 to 0.92) | 0.687 | (0.57 to 0.83) | .045* |
| Q3 | 0.490 | (0.40 to 0.61) | 0.520 | (0.42 to 0.65) | 0.420 | (0.32 to 0.55) | .007** |
| Q4 | 0.785 | (0.71 to 0.89) | 0.723 | (0.64 to 0.84) | 0.873 | (0.79 to 0.97) | .004** |
| Q5 | 0.835 | (0.72 to 0.97) | 0.909 | (0.81 to 1.00) | 0.760 | (0.61 to 0.94) | .035* |
| Q6 | 0.473 | (0.15 to 0.88) | 0.394 | (0.09 to 0.78) | 0.586 | (0.20 to 1.00) | 0.052 |
| Q7 | 0.847 | (0.72 to 0.98) | 0.825 | (0.67 to 0.99) | 0.892 | (0.78 to 1.00) | .343 |
| Q8 | 0.759 | (0.65 to 0.90) | 0.685 | (0.58 to 0.82) | 0.652 | (0.52 to 0.82) | .564 |
| Q9 | 0.735 | (0.65 to 0.84) | 0.748 | (0.65 to 0.87) | 0.733 | (0.62 to 0.86) | .828 |
| Q10 | 0.862 | (0.74 to 0.99) | 0.901 | (0.80 to 1.00) | 0.829 | (0.60 to 1.00) | .117 |
| **WHOLE EXAM** | 0.710 | (0.67 to 0.77) | 0.722 | (0.69 to 0.78) | 0.698 | (0.66 to 0.76) | 0.224 |

**Table 2: Results of the analysis comparing the inter-reliability of the blind versus the visible condition**

Our analysis shows that the blind condition is significantly more reliable for exam questions 2, 3 and 5, whereas the visible condition is significantly more reliable for exam question 4. When calculating the overall exam $\kappa$ including all feedback items of the exam (weighted according to their score), the inter-rater reliability of the blind condition is not significantly different from the inter-rater reliability of the visible condition ($p$=.224).

## Discussion

A possible explanation for why blind grading outperformed visible grading in terms of inter-rater reliability for questions 2, 3 and 5, is the strictness of the correction scheme. For example, in question 3, one checkbox could only be selected if a list of keywords was included in the student's answer (see checkbox 'right explanation of $C_{11}$' in the exam available at Moons (2021), which was a very strict rule to follow. We see that almost all assessors obey this requirement in the blind condition. In contrast, the assessors in the visible condition, more aware of the impact of not checking the item on the final grade, are less strict and check the box more quickly when the wording is somehow okay, even when some keywords are missing. Similar considerations have probably been given in question 5 (see Figure 3): the checkbox 'curly bracket is missing' is used much less frequently in the visible grading condition, even though they are assessing the same students. When the student's answer resembled a linear equation system, it was more often assessed as fine in the visible condition. Assessors in the blind condition had fewer reservations about ticking the item as they did not know the student would lose 1/3 of the points on this question by checking the box. In question 4, visible grading exhibits significantly higher inter-rater reliability. Based on an analysis of the assessors' judgements, this is likely to be related to the relative complexity of the correction scheme for this question. As assessors see the grade they are giving, they can easily see if their correction is likely to be correct when the same grade is given to a student with a similar answer. Indeed, in the visible condition, scores also function as a feedback mechanism to the assessor. If assessors obtain a similar score for a similar student answer, they probably assume their assessment is correct. In the blind condition, this mechanism is lost. Participating assessors also addressed this loss in a survey conducted right after their correction work (Moons & Vandervieren, 2022).

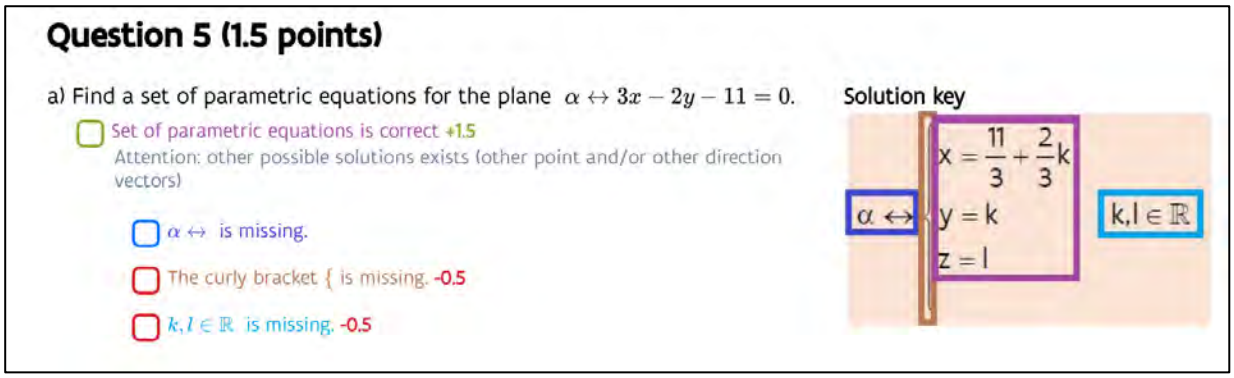


**Figure 2: Question 5 of the mathematics exam**

Referring back to the DiaCoM framework (see Figure 2), it seems that the *diagnostic behaviour* from assessors is influenced by their *diagnostic thinking;* more specifically, their *interpretation* of the value of partial scores seems to be an important factor on how the checkbox items are *perceived,* which then influences their decision making (too harsh/lenient). Indeed, as conclusion, we can say blind grading enhances inter-rater reliability when the correction scheme is very strict in what is correct or not. Strictness is defined as leaving no room to value answers that are nearly correct. Moreover, the correction scheme should not be too complex, otherwise visible grading is to be preferred for the feedback-loop visible partial scores provide. Although *personal* characteristics were not investigated in this paper, we know that assessors like checkbox grading in general, but always prefer visible over blind checkbox grading (Moons & Vandervieren, 2022): they perceive blind checkbox grading as less useful, less easier to use and more terrifying to use than visible checkbox grading.

The fact that we did not get a clearer picture on the inter-rater reliability is also linked to the limitation of the study that the assessors in the Flemish Exam Commission do not know the students they are assessing; as such, they are less prone to biases (Baird et al., 2004). A replication of the study in a standard classroom setting with teachers as assessors of their own students, could yield more convincing results in favour of blind grading.

In future research, we will also explore the students' perception of the feedback from the checkbox grading system and the assessors' behaviour and perception while using the system.

## 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

Baird, J., Greatorex, J., & Bell, J. F. (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

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.*

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

Dawson, P. (2017). Assessment rubrics: Towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education*, *42*(3), 347–360. https://doi.org/10.1080/02602938.2015.1111294

Doğan, C. D., & Uluman, M. (2017). A comparison of rubrics and graded category rating scales with various methods regarding raters' reliability. *Educational Sciences: Theory & Practice*, *17*, 631–651. https://doi.org/10.12738/estp.2017.2.0321

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. (2021). A tool for comparing mathematics tasks from paper-based and digital environments. *International Journal of Science and Mathematics Education*, *19*, 1655–1675. https://doi.org/10.1007/s10763-020-10119-0

Loibl, K., Leuders, T., & Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (DiaCoM). *Teaching and Teacher Education*, *91*, 103059. https://doi.org/10.1016/j.tate.2020.103059

Moons, F. (2021). The handwritten questions and their checkbox grading schemes of the mathematics exam of the study. https://mathsa.uantwerpen.be/ExamEnglish.pdf

Moons, F. & Vandervieren, E. (2022). Handwritten math exams with multiple assessors: Researching the added value of semi-automated assessment with atomic feedback. In J. Hodgen, E. Geraniou, G. Bolondi, & F. Ferretti (Eds.), *Proceedings of the Twelfth Congress of European Research in Mathematics Education (CERME12)*. (pp. 3859–3866). ERME / Free University of Bozen-Bolzano. https://hal.science/hal-03753446

Moons, F. & Vandervieren, E. (2023). Measuring agreement among several raters classifying subjects into one-or-more (hierarchical) nominal categories. A generalisation of Fleiss' kappa. *arXiv preprint*. https://doi.org/10.48550/arXiv.2303.12502

Moons, F., Vandervieren, E., & Colpaert, J. (2022). Atomic, reusable feedback: A semi-automated solution for assessing handwritten tasks? A crossover experiment with mathematics teachers. *Computers & Education Open, 3*, 100086. https://doi.org/10.1016/j.caeo.2022.100086

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