ZDM – Mathematics Education
https://doi.org/10.1007/s11858-024-01550-6

ORIGINAL PAPER

Checkbox grading of handwritten mathematics exams with multiple assessors: how do students react to the resulting atomic feedback? A mixed-method study

Filip Moons[1,2] · **Paola Iannone**[3] · **Ellen Vandervieren**[2]

Accepted: 18 January 2024
© The Author(s) 2024

Abstract
Handwritten tasks are better suited than digital ones to assess higher-order mathematics skills, as students can express themselves more freely. However, maintaining reliability and providing feedback can be challenging when assessing high-stakes, handwritten mathematics exams involving multiple assessors. This paper discusses a new semi-automated grading approach called 'checkbox grading'. Checkbox grading gives each assessor a list of checkboxes consisting of feedback items for each task. The assessor then ticks those feedback items which apply to the student's solution. Dependencies between the checkboxes can be set to ensure all assessors take the same route on the grading scheme. The system then automatically calculates the grade and provides atomic feedback to the student, giving a detailed insight into what went wrong and how the grade was obtained. Atomic feedback consists of a set of format requirements for mathematical feedback items, which has been shown to increase feedback's reusability. Checkbox grading was tested during the final high school mathematics exam (grade 12) organised by the Flemish Exam Commission, with 60 students and 10 assessors. This paper focuses on students' perceptions of the received checkbox grading feedback and how easily they interpreted it. After the exam was graded, all students were sent an online questionnaire, including their personalised exam feedback. The questionnaire was filled in by 36 students, and 4 of them participated in semi-structured interviews. Findings suggest that students could interpret the feedback from checkbox grading well, with no correlation between students' exam scores and feedback understanding. Therefore, we suggest that checkbox grading is an effective way to provide feedback, also for students with shaky subject matter knowledge.

Keywords Feedback · Atomic feedback · Student responses · Receptivity · Engagement with feedback · Cognitive processing · Computer-assisted assessment · State examinations · Summative feedback

1 Introduction

Assessing handwritten tasks is often a daunting endeavour in large-scale exams where multiple assessors are involved. Finding efficient ways to provide (consistent) feedback and reliable grading in these settings is not straightforward (Baird et al., 2004; Meadows & Billington, 2005; Morgan & Watson, 2002). *Grading reliability* is the degree to which a grade genuinely reflects the quality of a student's assignment, in which aspects outside the assignment should not play any role and is most often measured by letting multiple assessors rate the same assignment (Bloxham et al., 2016). Most exam designers try to ensure inter-rater reliability by pre-developing a solution key with grading criteria for assessors (Ahmed & Pollitt, 2011). Traditionally, these grading instructions also serve as a feedback tool, as students should be able to trace back the grades they obtained by comparing their solutions with the grading criteria (Price et al., 2012). Nevertheless, students often struggle to understand the language used in grading criteria (Cartney, 2010).

In an attempt to tackle the challenges of assessing handwritten mathematics tasks, we invented 'checkbox grading'. It involves a semi-automated assessment method

✉ Filip Moons
f.moons@uu.nl

[1] Freudenthal Institute, Utrecht University, Utrecht, The Netherlands

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

[3] School of Mathematics, The University of Edinburgh, Edinburgh, UK

Published online: 09 March 2024

(Moons et al., 2022) in which assessors work with a software tool to grade and provide feedback on handwritten mathematics tasks. The method was researched during the final high school mathematics exam (grade 12) organised by the Flemish Exam Commission. The method can possibly make the grading process more efficient and reliable. Moreover, the grading method results in *atomic* feedback, giving a detailed insight into how grades were obtained. Atomic feedback consists of a hierarchical list of feedback items, resulting in a separate feedback item for each independent error.

Why is a semi-automated method still necessary? Can all the difficulties caused by assessing handwritten tasks be avoided by replacing them with digital exams that can be assessed fully-automatedly (Kloosterman & Warren, 2014)? Many studies have concluded that some mathematical learning goals lend themselves well to being assessed digitally—other, much less. For example, Hoogland and Tout (2018) warn that digital mathematics tasks often focus on lower-order goals (e.g., procedural skills). They argue that handwritten tasks better assess vital higher-order goals (e.g., problem-solving skills). Bokhove and Drijvers (2010) point out that handwritten tasks allow students to express themselves more freely. Lemmo (2021) highlights substantial differences in students' thinking processes when the same task is asked digitally or paper-based. Together, these studies emphasise the continued significance of handwritten mathematics exam tasks. When designing a mathematics exam, it is best to decide individually whether the digital or handwritten mode is appropriate for each task, leading to exams that are a mixture of both (Threlfall et al., 2007).

This paper investigates the student perspective, specifically exploring whether students can effectively interpret the feedback provided through checkbox grading. Additionally, we have discussed the assessors' usage and perspectives separately in another paper (Moons et al., 2023b) and presented a statistical method for measuring grading reliability resulting from this approach (Moons & Vandervieren, 2023a).

1.1 Checkbox grading

1.1.1 Idea

Using checkbox grading, exam designers produce a grading scheme for each task consisting of different feedback items written in an atomic way (see Sect. 1.1.2.), anticipating the mistakes students may make in the given question. Next, students solve these exam tasks the classical way by writing on paper. Subsequently, the papers are scanned, and the assessors use the checkbox grading system to assess the solutions on a computer. When correcting a student's solution, the assessors just select the feedback items ('checkboxes') that apply.

To allocate grades, exam designers can associate items with partial points to be added (green items in Fig. 1) or subtracted (red items in Fig. 1). It is also possible to associate items with a threshold (e.g., 'if this feedback item is ticked, maximum 1 out of 2 points'). Items that do not change the grade but provide essential information for the continuation of the assessment have a blue checkbox (e.g., as a note to the assessors that some solutions are fine as well or as an indicator for the system to know how to proceed). 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 items can be set so that items can be shown, disabled, or changed whenever a previous item is ticked, implying that the assessors must follow the point-by-point list from top to bottom. This adaptive sequencing is the main characteristic of the approach, resembling a flow chart that automatically determines the grade. By ticking the checkboxes that are relevant to a student's answer, we envision (1) a deep insight into how the grade was obtained for both the student (feedback) as well as the Flemish exam commission, and (2) a straightforward way to grade handwritten mathematics tasks when multiple assessors are involved.

An example of this approach is given in Fig. 1. The exam task consists of two sub-tasks. In the first sub-task, the student makes a mistake on the sign. As the item 'Answer is completely correct' is unticked, the computer knows that a mistake happened; therefore, the system shows two additional blue checkboxes to decide whether the assessor can continue grading the task. The sign error was an anticipated mistake that caused a deviation from the solution key. While the student did not gain points with sub-task (a), the assessor might continue with the assessment but now has to check the solution individually by calculating along. Any other mistake in sub-task (a) would have stopped the further assessment of the task. In sub-task (b), the student corrects the previous mistake but fails to provide the correct solution. As such, only the first item of sub-task (b) ('The row echelon form is correct') applies, leading to a total score of 1/3.5.

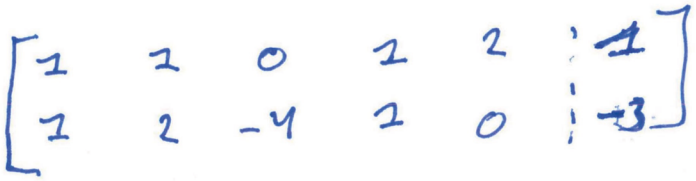
If a particular solution approach by a student is not covered in the available feedback items, an assessor could add additional checkboxes. Checkbox grading was developed as an advanced grading method plug-in for Moodle, an open-source e-learning platform (Gamage et al., 2022).

Finally, the name 'checkbox grading' was inspired by the bestseller 'The Checklist Manifesto' (Gawande, 2010), in which the author argues that using simple checklists in daily and professional life can make even very complex processes efficient, consistent and safe: "under conditions of complexity, not only are checklists a help, they are required for success" (p. 79).

[4.5 points] Consider the following system of equations: $\begin{cases} x_1 + x_2 + x_4 + 2x_5 + 1 = 0 \\ x_1 + 2x_2 - 4x_3 + x_4 - 3 = 0 \end{cases}$

a) Write down the corresponding extendend coefficient matrix. (1 point)

Student's answer



Solution key

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

Correction by assessor (0/1)

☐ Answer is completely correct **+1.0**
- Answer also ok when | is missing, but the elements of the matrix are correct.

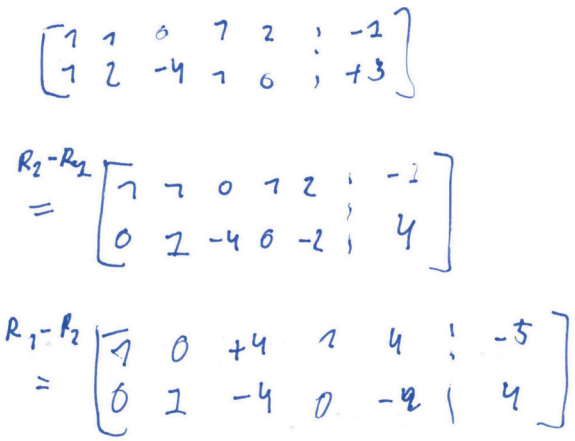
Check-up to see if you need to check the student's calculation individually...

☐ Answer is $\left[\begin{array}{ccccc|c} 1 & 1 & 0 & 1 & 2 & 1 \\ 1 & 2 & 4 & 1 & 0 & -3 \end{array}\right]$, check the students's calculation individually for subquestion (b)

☐ Answer is something different: no points for the rest of this question

b) Solve the system of equations: write down the row echelon form and the solution set. (2.5 points)

Student's answer



Solution key

$$\left[\begin{array}{ccccc|c} 1 & 0 & 4 & 1 & 4 & -5 \\ 0 & 1 & -4 & 0 & -2 & 4 \end{array}\right]$$

$$V = \{(-5 - 4k - l - 4m,\ 4 + 4k + 2m,\ k,\ l,\ m) \mid k,l,m \in \mathbb{R}\}$$

Correction by assessor (1/2.5)

☑ Check individually: The row echelon form is correct. **+1.0**

☐ The solutions x_1, x_2, x_3, x_4, x_5 were calculated correctly. **+1.5**

☐ No quintuples were written down because the brackets are missing. **max: 1.0**

☐ sol S =, V =, OV= or the curly braces {} are missing. **-0.5**

Grade: 1/3.5

Fig. 1 Checkbox grading scheme of exam task 4

1.1.2 Atomic feedback

The feedback in checkbox grading is called *atomic* feedback (Moons et al., 2022, 2024). Classic written feedback has traditionally consisted of long pieces of written text (Winstone et al., 2017). With its long sentences describing all the errors in a student's work, classic written feedback is intrinsically not reusable, as it is too explicitly targeted toward specific students. To overcome this difficulty and maximise the reusability of feedback, one of the key ideas underlying the checkbox grading system is that it encourages exam designers to write atomic feedback by breaking it into separate feedback items. To do so, one must (1) identify the possible independent errors occurring and (2) write separate feedback items for each error, independent of each other (making them atomic). 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 the indentation in Fig. 1). An additional criterion for being atomic holds for checkbox grading: (3) a knowledgeable assessor must be able to determine unambiguously whether an item applies to a student's answer. 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 colour to make their connection visually clear (see Fig. 1).

2 Theoretical background

In this section, we prepare for the study's research questions (Sect. 2.3) on the students' perspective on the received feedback through a short literature review on feedback interventions (Sect. 2.1) and the study's theoretical framework (Sect. 2.2).

2.1 Literature review

The efficacy of feedback interventions is ultimately determined by the degree to which students engage with the feedback content (Jonsson & Panadero, 2018). This engagement, termed '*proactive recipience*', is predicated on students' understanding of their feedback; without comprehension, feedback fails to facilitate improvement (Jonsson, 2013). In this research context, where 'checkbox grading' feedback is given after a high-stakes mathematics exam, we speak of *assessment literacy:* students' ability to comprehend and utilise the grading process to evaluate their performance (Winstone et al., 2017). In this context, assessment literacy is a prerequisite for proactive recipience, which can be supported by (1) understanding the connection between assessment, learning, and expectations, (2) assessing their own and others' performance based on specific criteria, (3) grasping the terminology and concepts used in feedback, and (4) becoming familiar with assessment methods and feedback practices (Price et al., 2012). Facilitating proactive recipience is especially important when the exam needs to be retaken. In addition, the provided checkbox grading feedback also aims to promote transparency so that students perceive the assessment as fair (Darabi Bazvand & Rasooli, 2022).

Several studies have investigated how engagement with grading criteria affects students' assessment literacy. Students generally rate these interventions positively (Atkinson & Lim, 2013) and see their importance (Orsmond et al., 2002), and some studies have shown that such interventions can improve grades and self-reported awareness of learning objectives (Case, 2007). Engaging with grading criteria seems to help learners understand the assessment process and expectations (O'Donovan et al., 2004; Rust et al., 2003). However, not all learners respond positively to these interventions (Bloxham & West, 2007), and some struggle to understand the language used in the grading criteria (Cartney, 2010). Additionally, understanding the grading criteria does not automatically translate to better future work (Rust et al., 2003).

2.2 Theoretical framework

2.2.1 Revised student-feedback interaction model

In 2016, Lipnevich et al. (2016) proposed a student-feedback interaction model (Fig. 2) that may be useful in considering the complexity of feedback and the factors that may affect student perceptions and subsequent action (or lack thereof). Later, Lipnevich and Smith (2022) revised the model, including a step-wise understanding of the feedback process. The model is based on several studies and meta-analyses on feedback and gives an overview of all the factors that relate to how students respond and interact with feedback. We will use this revised model to frame our research. The model suggests that feedback is received in a context that can influence how important or familiar the students perceive it. The interaction process starts with the feedback message and the source that generated it. Feedback can vary in tone, length, specificity, and complexity, and the source's trustworthiness plays an important role. Next, the model investigates how the student receives the feedback and how it is processed: cognitively, affectively, and behaviourally. Three main questions describe this student's feedback processing: Do I understand the feedback? How do I feel about the feedback? What am I going to do with the feedback? Answers to these questions provide the student with self-feedback (Panadero et al., 2019). The final step concerns actions, outcomes, and the growth that results from the feedback. The interaction model served as a guiding

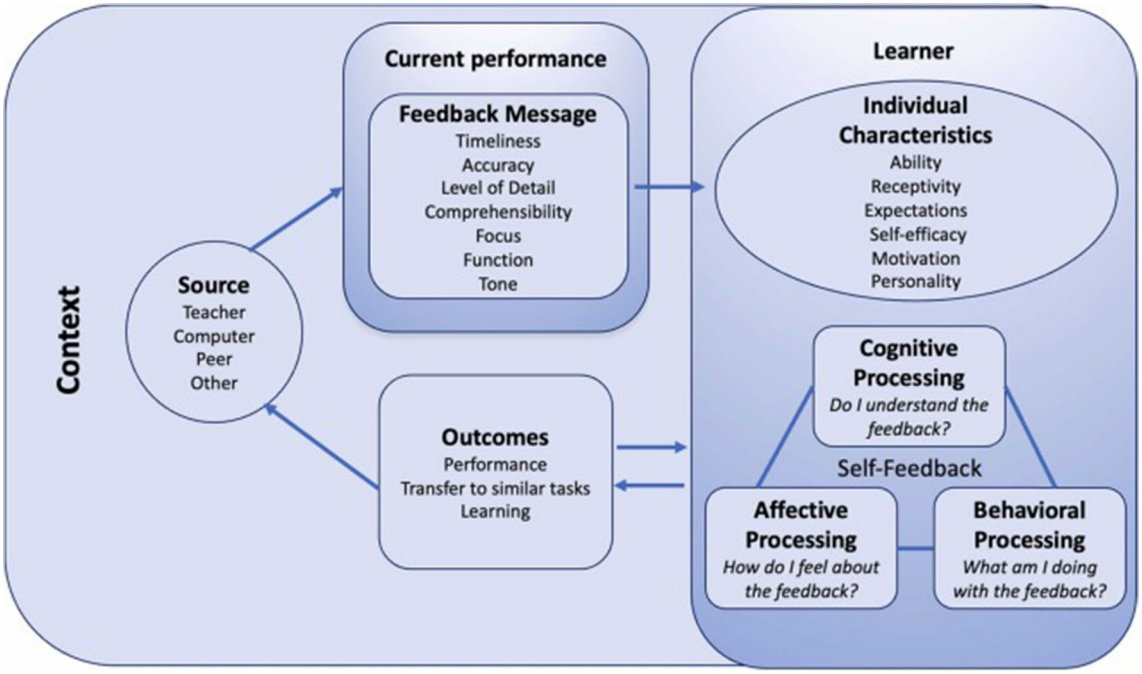


Fig. 2 The revised student-feedback interaction model by Lipnevich and Smith (2022)

theoretical framework for our study, as explained in the following section.

2.2.2 The student-feedback interaction model applied to our study

This study explores how students interpret and perceive feedback reports from checkbox grading. One approach to achieving this goal involved contrasting the checkbox grading feedback with various other delivery methods, such as classic written feedback, only communicating a grade and the Flemish Exam Commission's traditional grading procedure (see Fig. 5 to compare these feedback types). The traditional grading scheme of task 2 is shown in Fig. 3. In the traditional grading process, the assessors apply the provided grading scheme of a task to the student's solution and only communicate the grade they deduced. During a review appointment, students receive their exams, the traditional grading schemes, and their obtained grades. In doing so, students sometimes have to guess which criteria were applied to arrive at a particular grade (Price et al., 2012).

To connect our conditions to the revised student-feedback interaction model, the *context* consists of students taking a high-stakes mathematics exam to graduate from Flemish secondary education, a stressful and relatively uncommon context for most students. The *source* of the feedback is the Flemish Exam Commission. The *feedback message* is mainly presented as checkbox grading feedback, but some messages are also phrased using classic feedback approaches in order to compare them to checkbox grading feedback. We gather most *individual characteristics* through a questionnaire, as well as glimpses of the *cognitive* and *affective processing*. Semi-structured interviews were conducted to gain deeper insight into the *cognitive processing*. A blind spot in our study remains the *behavioural processing* and the resulting *outcomes*, as we did not follow the students who failed the exam on a second attempt.

2.3 Research questions

Now that we have established the theoretical and conceptual underpinnings of the study, we pose two research questions to guide our inquiry:

> [RQ 1] How do students understand (*cognitive processing*) and perceive (*affective processing*) feedback reports from checkbox grading?
> [RQ 2] How do students' perceptions of receiving feedback from checkbox grading differ from perceptions of feedback from classical approaches (such as traditional grading, written feedback, or communicating grades)?

Note that our prior aim is to investigate how students interpret checkbox grading feedback (RQ1); the other classical feedback approaches are included for comparison purposes (RQ2) but will not be investigated in depth.

2) (/2,5) Let: $z_1 = b.(\cos\alpha + i.\sin\alpha)$ and $z_2 = c.(\cos\beta + i.\sin\beta)$, with b,c $\in \mathbb{R}_0^+$.

Calculate the following expressions and write the answer in polar form.

a) $-5.z_1$

$-5 = 5.(\cos 180° + i.\sin 180°) \Rightarrow$

$-5.z_1 = 5b.(\cos(\alpha + 180°) + i.\sin(\alpha + 180°))$

OR:

$-5 = 5.(\cos\pi + i.\sin\pi) \Rightarrow$

$-5.z_1 = 5b.(\cos(\alpha + \pi) + i.\sin(\alpha + \pi))$

- 0,5 point for correctly converting -5 to polar form
  NOTE: May be combined with the next intermediate step
- 1 point for correct modulus and argument (modulus must be positive!)
  NOTE: -0.5 point when it was **converted back** to $-5b.(\cos\alpha + i.\sin\alpha)$, 0/1.5 if only $-5b.(\cos\alpha + i.\sin\alpha)$ was written down.

NOTE: -0,5 point if the brackets around the argument and/or around cos...+i.sin... are missing: **only apply when the maximum score was obtained** (1,5/1,5)

b) $z_1.z_2^3$

$z_1.z_2^3 = b.c^3.(\cos(\alpha + 3\beta) + i.\sin(\alpha + 3\beta))$

- 0,5 point (or 0) for correct modulus
- 0,5 point (or 0) for correct argument

NOTE:: 0,5/1 for $b.c^3$.(cos α+i sin α).(cos3β+i sin3β)

NOTE: -0,5 point if completely correct but the brackets around the argument and/or around cos...+i.sin... are missing, **unless already penalised in sub-task 2a.**

Fig. 3 Traditional grading scheme of exam task 2

3 Methods

The study was conducted with the Exam Commission of Flanders (the Dutch-speaking part of Belgium) for Secondary Education. Flanders is a region without any central exams (Bolondi et al., 2019): every secondary school decides autonomously on the assessment of students. Consequently, the Exam Commission does not organise national exams for all Flemish students. However, they organise large-scale exams for anyone who cannot graduate from the regular school system. In this way, students who pass all their exams with the Exam Commission can still obtain a secondary education diploma. Students participating in these exams prepare by themselves or with the support of a private tutor/school. The commission provides clear guidelines for students on the content of the exams, carries out the exams, and awards diplomas but does not provide any teaching activities or materials to students.

Ethical clearance for this study was obtained from the University of Antwerp Ethics Committee. The Committee approved the study design and the procedures for data management, consent, and protecting the participants' privacy.

In the following sections, we describe the mathematics exam used in this study, the development of the questionnaire and interview protocols, the sample of participating students and the data analysis procedures.

3.1 Mathematics exam

The mathematics exam for this study was developed by the three mathematics exam designers of the Flemish Exam Commission (without any influence from the researchers) following their standard practice. The exam is part of the advanced mathematics track of Flemish secondary education's senior years (11th/12th grade). The exam is a mixture of digital and paper-and-pencil tasks: 46% of the exam grades are obtained with the digital part and 54% with paper-and-pencil tasks. In this study, only the feedback on the paper-and-pencil tasks is considered. These tasks vary considerably in points allocated based on the importance of the topic in the curriculum; 0.5 points was the smallest partial score. An overview of the content of the exam can be found in Table 1.

A timeline of the study can be found in Fig. 4. The study started when the exam designers had prepared the exam by the 1st of October 2021. Next, their traditional solution key with grading instructions was turned into checkbox grading in close cooperation with the researchers. After the exam, the assessors (mostly mathematics teachers who do this as

Table 1 Content of the mathematics exam, including the maximum, mean and standard deviation of the scores of the students who filled in the questionnaire

#	Topic	Learning goal	Max score	Avg. score M ± SD
Paper-based tasks			**54**	**17.1 ± 9.8**
T1	Complex numbers	Calculations with complex numbers in $a + bi$-form	2.5	1.7 ± 1.0
T2	Complex numbers	Calculations with complex numbers in polar form	2.5	0.6 ± 0.6
T3	Matrices	Modelling with matrices	3.5	1.9 ± 1.1
T4	Matrices	Coefficient matrices of linear equations	3.5	1.2 ± 1.1
T5	Solid geometry	Parameter equations of a plane	1.5	0.1 ± 0.4
T6	Solid geometry	Cartesian equation of a line	1	0.1 ± 0.2
T7	Solid geometry	Drawing a segment line in the x, y, z-axis system	2.5	1.2 ± 0.8
T8	Solid geometry	Determining the distance between a point and a line	4.5	0.6 ± 1.3
T9	Solid geometry	Parallel lines in solid geometry	2.5	0.9 ± 0.9
T10	Probability	Modelling a probability experiment	4	0.5 ± 1.1
Digital part			**46**	**20.4 ± 8.5**
Algebra			3	1.8 ± 1.1
Solid geometry			6	3.9 ± 2.2
Discrete mathematics			11	2.6 ± 2.3
Statistics			21	10.7 ± 4.6
Research competencies			5	1.5 ± 1.8
Total			**100**	**37.5 ± 16.3**

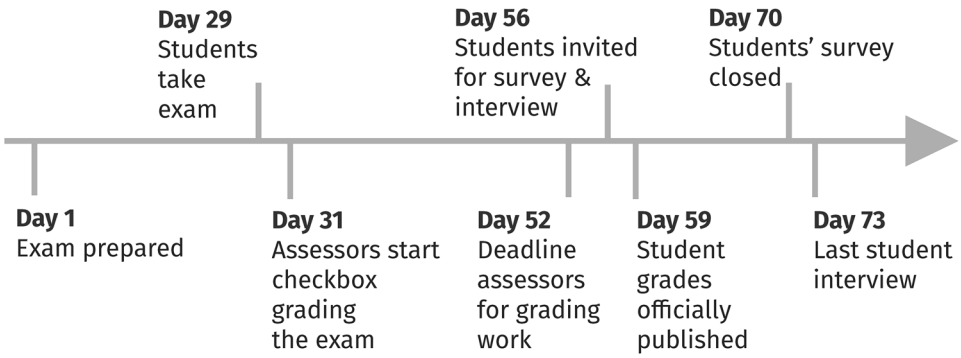


Fig. 4 Timeline of the study

a side job for the Exam Commission) used the checkbox grading system to assess the exam. All the paper-based tasks of the exam, including the checkbox grading schemes, can be found in the electronic supplementary material.

3.2 Instrument development

3.2.1 Questionnaire

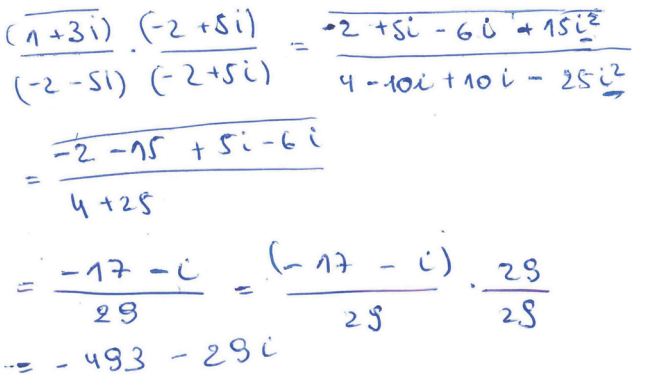
The questionnaire was implemented in Qualtrics, consisted of four parts and was developed based on our presented theoretical framework (Lipnevich & Smith, 2022). A key aspect was to keep the completion time below 15 min to motivate students to answer truthfully until the end (Yan et al., 2011). The four parts of the questionnaire were:

1. *Individual characteristics and past experiences.* The first part gathered some personal information about the students (age, study direction, reasons to opt for the Exam Commission, and number of previous exam attempts). We also asked how the students experience the current feedback practices at the Flemish Exam Commission.
2. *Ranking exercise on the comprehensibility of the feedback types (RQ2).* In the second part of the survey, students ranked four types of feedback from most comprehensible to least comprehensible by drag-and-drop. All feedback types dealt with the same exemplar task from a peer student, were content-wise equivalent and resulted in the same grade; the only difference was their appearance. The four feedback types were checkbox grading, classic written feedback, only a

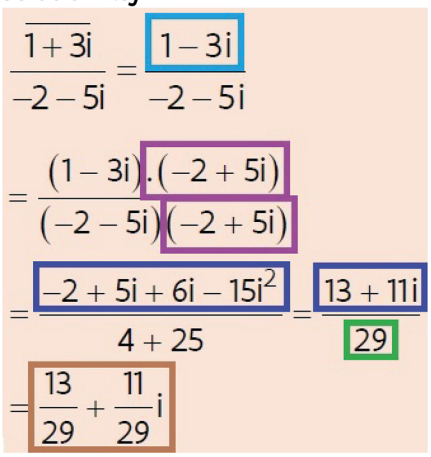
grade, and traditional grading. The four feedback types were adapted from Harks et al. (2014) and Koenka et al. (2019). This ranking question was repeated for two different exam tasks to avoid a dependency between the type of task and the preferred feedback. An example of one of the ranking questions can be found in Figure 5.

Here is an exam task and an answer a fellow student gave:

Calculate $\frac{\overline{1+3i}}{-2-5i}$ and write the answer in $a+bi$ form.
Show all your intermediate steps, don't use your calculator.

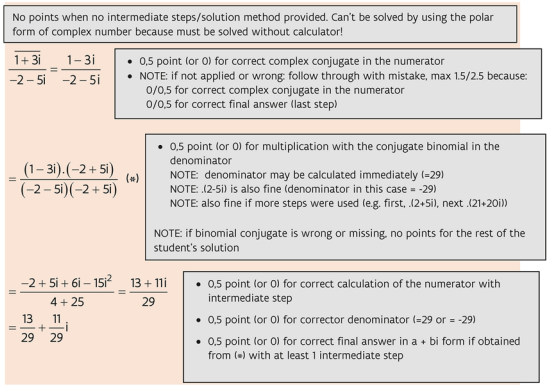


Solution key



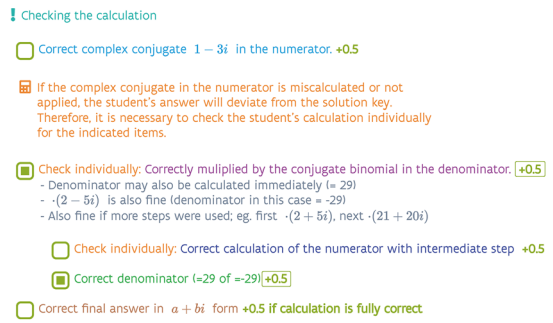
Rank the following four feedback types (that are equal in content and grade) from most to least comprehenible.

Traditional grading



Grade: 1/2.5

Checkbox grading



Grade: 1/2.5

Classic feedback

It is not clear if you can determine the complex conjugate of $1+3i$. You correctly multiply the numerator and denominator with the conjugate binomial $-2+5i$. The numerator is wrongly calculated, and it is unclear where the sign error comes from (an error in the complex conjugate or just a calculation mistake). The denominator is correctly determined (=29). The result is completely wrong because of the mistake with the numerator.

Grade: 1/2.5

Only a grade

Grade: 1/2.5

Fig. 5 Ranking exercise on the comprehensibility of different feedback types on the same solution of exam task 1

3. *Quiz on understanding the feedback given to a fellow student (RQ1).* In the third part, students saw the feedback report depicted in Figure 1. They were asked to answer 10 short false/true questions about the feedback content, as shown in Table 3. The questions polled their understanding of the feedback and the sequencing of the grading scheme. Students had to answer each question and could not return to previous questions as some following questions sometimes revealed the answer of a previous one.
4. *Personal checkbox grading feedback: student's cognitive & affective processing (RQ1).* In the last part, students received a link to access their personal checkbox feedback on exam tasks 1, 7 and 10, which looked like the feedback report in Figure 1. Based on Weaver (2006), we tried to measure how students perceived the personal checkbox grading feedback they received. The survey questions can be found in Figure 6.

3.2.2 Interview protocol

The semi-structured interviews of students took place online at most a week after completing the survey. The interviews investigated the students' understanding of their exam feedback. We used open questions and a think-aloud protocol (Gillham, 2005) to reveal their thinking while processing their feedback. One researcher prepared each interview by scanning the student's exams and indicating interesting solutions for exam tasks to discuss. The chosen exam tasks were usually (partially) incorrect, as these are the best triggers to see if students understand what should be improved. Correct exam tasks were occasionally discussed with hypothetical supplementary interview questions (e.g., 'What would have happened if your numerator had been wrong?'). When traditional grading was discussed, the students saw the traditional grading scheme of the task (Fig. 3) and their solution. When checkbox grading was discussed, they saw their complete feedback report (Fig. 1). The interview protocol contained two interview questions inspired by O'Donovan et al. (2004):

1. *Cognitive processing of traditional grading.* I'm sharing my screen, showing your solution to exam task x and the traditional grading scheme. Can you determine the grade you should receive and explain your reasoning?
2. *Cognitive processing of checkbox grading.* I'm showing you your feedback report on exam task x. Can you think aloud about how your grade was obtained? What was correct in your solution? What was wrong or missing?

Exam task x was always replaced with the task number the researcher had chosen in advance. During the interviews, exam task 2 was chosen for all students to investigate their interpretation of traditional grading, and three to four other tasks were chosen to investigate their interpretation of checkbox grading, as this was the main focus.

The researcher always let the students talk and intervened only: (1) to remind students to think aloud, (2) when clarifications of their reasoning were necessary, or (3) to ask a follow-up question when a student made an incorrect interpretation. Follow-up questions were formulated as open and non-corrective as possible. In the case of an incorrect interpretation, the researcher briefly summarised the student's conclusion as a first follow-up question (e.g., 'So you are saying that..?'). If a student did not correct themselves after hearing the researcher's summary of their incorrect interpretation, one more follow-up question was asked, such as 'But does that hold for your solution?'.

3.3 Participants

3.3.1 Sampling

When all assessors finished their work (see Fig. 4), the questionnaire was ready to be sent to all students. Students received personal checkbox grading feedback on three exam tasks during the questionnaire as an incentive. Furthermore, upon completing the questionnaire, students were sent their final exam score, which was sooner than the official announcement. The questionnaire was closed two weeks after its release.

At the end of the questionnaire, students were asked if they would like to take part in an in-depth online interview of 45 min about the feedback they received on their exam. As an incentive to participate in an interview, they would receive their personal checkbox grading feedback on the whole exam (not just three tasks), eliminating the need for a traditional review appointment in Brussels.

3.3.2 Questionnaire

The questionnaire was filled in by 36 of the 60 students who took the exam. In total, 19 female students and 17 male students participated. They were, on average, 17.39 years old ($SD = 1.46$). All the students had advanced mathematics as part of the curriculum of their studies. The exam results of the 36 students who participated in the questionnaire can be found in Table 1. The exam results were, on average, relatively low. Indeed, many students just come to an exam session to know what preparation they need for the following session. Exactly half of our participants took the exam for the first time, 14 for the second time and 6 for the third time.

3.3.3 Interviews

Four of the 36 students who filled in the questionnaire agreed to be interviewed: Sacha (female, 17 years old, scored 19%), Jana (female, 17 years old, scored 42%), Tom (male, 17 years old, scored 60%) and Emile (male, 19 years old, scored 41%).

3.4 Data analysis procedures

3.4.1 Questionnaire

The questionnaire analysis mainly consists of a descriptive analysis of the results. Additionally, the average ranks were calculated for the ranking exercise (part 2), and a correlation test was executed for the quiz on feedback understanding (part 3).

3.4.2 Interviews

In the preparatory stage, interviews were transcribed verbatim by the researchers. A straightforward 'traffic light coding' procedure was implemented for each exam task discussed during the interview, as shown in Table 2. Another researcher double-checked the coding.

Table 2 Traffic light coding procedure of the interviews. The *x* denotes the number of the exam task

x	The student could independently make a correct interpretation of the given feedback without any help from the researcher
x	The student could correctly interpret the given feedback when the researcher asked a maximum of two follow-up questions
x	The student incorrectly interpreted the given feedback

4 Results

4.1 Cognitive and affective processing of checkbox grading feedback [RQ1]

The quiz aimed to ascertain student's understanding of the feedback in Fig. 1. As 'understanding the given feedback' can be seen as a latent construct, we analysed the composite reliability (Brunner & Süb, 2005) of the 10 initial questions. Three questions were deleted to achieve an acceptable composite reliability of 0.72. It seemed that these deleted questions could be interpreted ambiguously. The 7 remaining questions and the results can be found in Table 3. It shows that, overall, students understood how the checkbox grading is used in the exam they just took: on average, the students scored 72.6% (*SD*: 18.2%) on the quiz, which is much higher than the mean score of 37.5% on the actual exam (*SD:* 16.3%, see Table 1). A Pearson correlation coefficient was computed between students' quiz and exam scores, which showed no correlation between the two vari ables, $r(34) = -0.02$, $p = 0.91$ with 95% CI [−0.34, 0.31].

Moreover, the results of the last part of the survey, in which students could access their personal checkbox grading feedback on three questions, can be found in Fig. 6. The results on students' understanding and affective processing indicate that they would greatly appreciate it if the Exam Commission would adopt this approach. Students feel that they understand their feedback, learn from it, and see the connection with the grades they obtained.

These results are confirmed by analysis of the interviews (see Table 4). Table 4 shows that students could independently draw correct conclusions in 11 of the 16 exam tasks discussed, confirming the survey results. Below we summarise the interpretations the students gave of the feedback form the checkbox grading.

Sasha could interpret the checkbox grading feedback she received on 3 of the 4 tasks. She scored 1.5/2.5 on task 1 by not writing one intermediate step explaining her

Table 3 Quiz on the understanding of the checkbox grading feedback shown in Fig. 1

#	Item	Correct answer	% correct
1	The student's extended coefficient matrix in sub-task a is correct	False	72.2
2	The student's extended coefficient matrix in sub-task a is wrong but 'good enough' to continue the assessment, taking into account the mistake	True	77.8
3	If the extended coefficient matrix had contained other mistakes in sub-task (a), then no points could be awarded for sub-task (b)	True	69.4
4	The student's row echelon form of the student in sub-task (b), is effectively the form you should have obtained	True	69.4
5	The student gets only 1 point for sub-task (b) because no solution set was written down	True	88.3
6	The student gets only 1 point for sub-task (b) because brackets do not enclose the quintuples	False	50.0
7	If the student had written down a completely correct set, the total of this task would have been 2.5/3.5	True	86.1

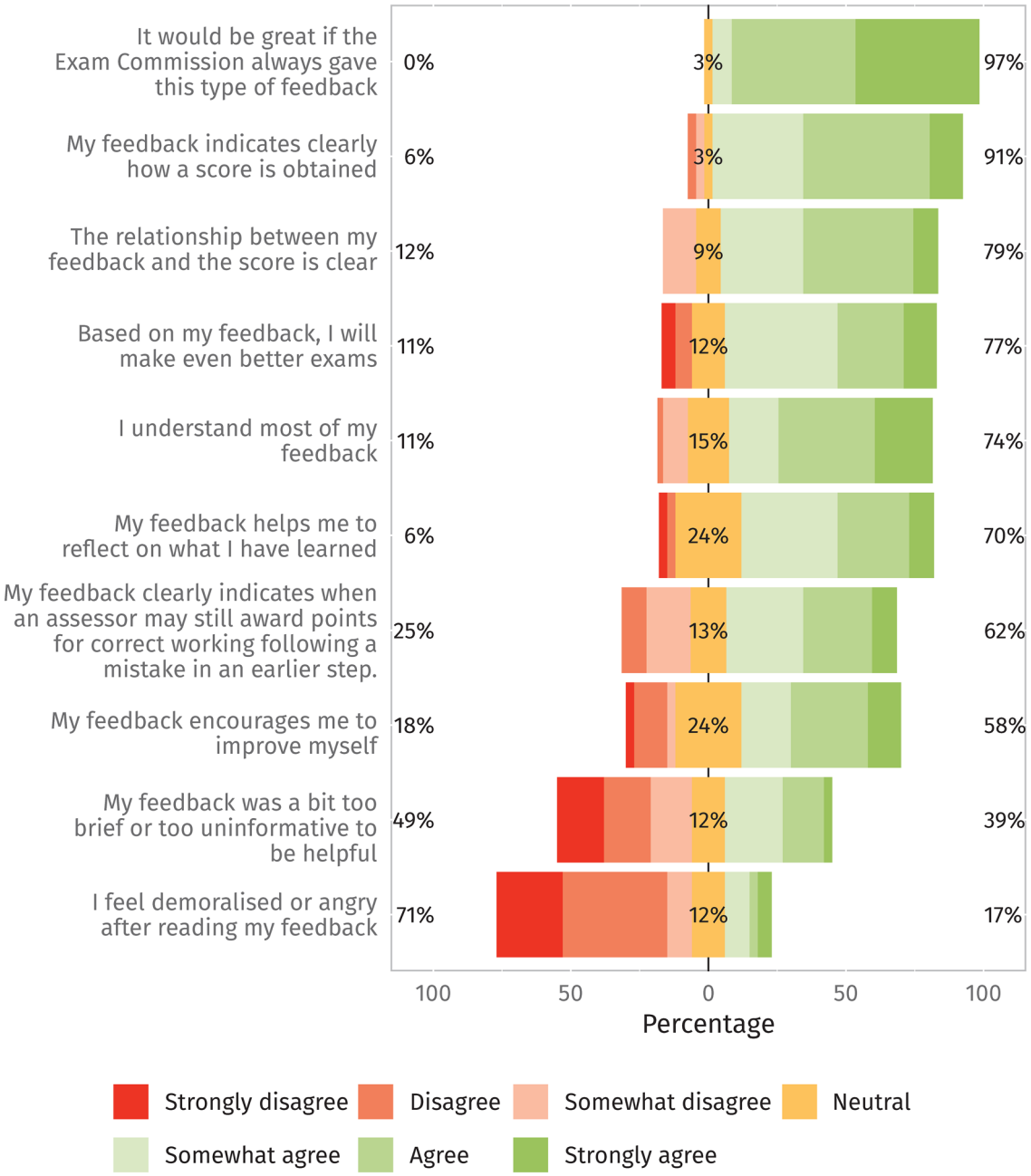
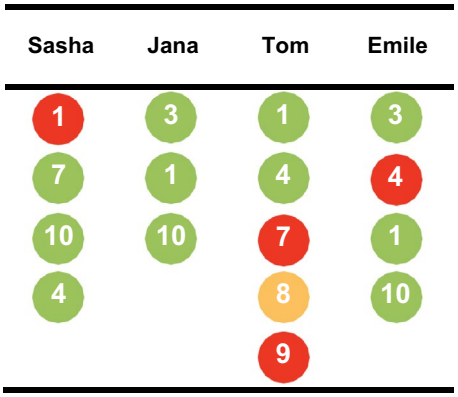


Fig. 6 Overview of the students' survey items corresponding to their checkbox feedback

calculation. At the same time, the instructions indicated that all intermediate steps should be shown as no calculator could be used (see Fig. 5). She worried when seeing the first item that would lead to a zero score on task 1: 'No intermediate steps provided' (which did not apply to her solution). When the researcher asked if the item applied to her solution (although the box was not checked), she insisted it did. She probably confused it with the only intermediate step missing in her solution, for which she was indeed penalised by 0.5 points as the box 'Correct calculation of the numerator with intermediate step' was not checked. When asked why her final score for task 1 was 1.5/2.5 and not 2/2.5, she said the indentation of this checkbox was probably the reason. It was not: the indentation indicated a parent–child sequence in the grading scheme: this (child) item could only be selected when the parent item was. The researcher then explained that the missed 0.5 points came from the last item, which an assessor could select but only added + 0.5 when everything else was fully correct. After this clarification, Sasha correctly interpreted the checkbox grading feedback

Table 4 Results of the traffic light analysis of the cognitive processing of checkbox grading



on all questions that followed. Even hypothetical questions like 'If the assessor would have selected the item that you drew the segment line $[AB]$ correctly, would it have changed something to your score?' (task 7, answer: no) were answered correctly. Moreover, she could independently say what she has to change in her solutions when taking the exam a second time. To conclude, Sasha showed a high assessment literacy with checkbox grading, which was somewhat surprising due to her low final score on this second exam (19%).

Jana could independently interpret her checkbox grading on all three tasks discussed in the interview. She correctly inferred her result on task 3 (3/3.5), and could indicate that she missed some of the keywords needed for obtaining full marks. She repeatedly stressed how clear she found the checkbox grading scheme, for example, when discussing task 1 (see Fig. 5):

> "For the first step, it is already very clearly stated that $1-3i$ must be present in light blue colour, and this is also present in the solution key in light blue, so they are clearly connected. You immediately know which expressions are linked together. And it is really step-by-step: in this step, this must be present; in this step, you get these points. This and the link with the colours: very clever. This is so much more clear (than traditional grading, ed.) (…) If you got it wrong, you could say: here I was wrong, and my solution is not correct anymore because I made a mistake here." (Jana).

Tom started the interview by stressing that he liked the checkbox grading much more than the traditional grading schemes. The discussion began with his correct solution to task 1 (2.5/2.5). On the hypothetical question of what his score would have been if the numerator had been wrong, after encouragement to read the entire checkbox grading scheme, he correctly concluded it would be 1.5/2.5 because the last item only adds + 0.5 if everything else was correct. For task 4 (2/3.5), he immediately indicated that his solution was missing some elements. Interestingly, when the researcher was scrolling through his exam paper, Tom asked to discuss tasks 7 and 8. For task 7 (1/2.5), he said he could not correctly draw the point B as he forgot to bring a set square, implicitly indicating he understood the feedback. When the researcher asked what would have happened to the grade if he had drawn the segment, Tom replied he would have received + 0.5 extra points. This interpretation was incorrect: even if the assessors had selected the item, it would not have changed the grade as the item indicates that + 0.5 is only awarded when points A and B are drawn correctly. Hence, Tom struggled to understand the sequencing in the grading scheme. In task 8 (4/4.5), his solution to the task was correct, but he lost half a point because he needed to explain his reasoning. Tom said he did not understand why he lost half a point for this and made it clear that he disagreed with the grading criteria. When the researcher asked whether it is important to justify the steps in mathematical reasoning, Tom reluctantly agreed. This exchange was coded in Table 4 as orange because it is likely that Tom understood the feedback after the remark from the interviewer, although he disagreed with it. Finally, in task 9 (2/2.5), a remark was added by the assessors pointing to an inappropriate use of double arrows, which makes for a half-point loss. While Tom understood the mechanism behind checkbox grading, knowing that selecting such an additional remark affected his grade by -0.5, he said:

> "Yes, but I do not understand it. I've read it, but I don't understand it. Why is there a problem with the double arrows?" (Tom)

We could only conclude that this additional remark was too short for Tom; therefore, this task was coded in red in Table 4. Interestingly, this was the first time during the interviews that a lack of content knowledge was the cause of a lack of understanding of the feedback.

Finally, Emile, who only needed 2% extra to pass, could interpret the feedback on his task 3 (2.5/3.5) well but reacted emotionally to the feedback on sub-task 3(b):

> "Ow ow ow ow. Oh, wait. Wait, I don't have that?! Oooooh, I have written 0.013 and not 1.013. I had to add 1! Oooooooh nooooooo. Oooh, such a bummer! That would have been my needed 2%! That would have been my needed 2%!" (Emile)

When discussing sub-task 4(b) (2/3.5), he gave the solution $\{-1, 4\}$, which is impossible as there are 5 unknowns. His solution was far from correct, and Emile could not link the unchecked items to his solution.

"Well, I really don't know what I am doing there (with the solution set, ed.). I failed to solve that last part, and I also don't know how you should have solved it." (Emile)

While discussing his correct task 1 (2.5/2.5), Emile mentioned that he found this kind of feedback very clear because it is easy to see what contributed to the grade and how the feedback and solution are linked together due to the use of colours. Finally, for task 10 (0/4), he could correctly interpret the sequencing in the grading scheme.

4.2 Perceptions of feedback of checkbox grading compared to classical approaches [RQ2]

Regarding the second research question, past students' experiences were surveyed in the first part of the questionnaire. Most students (52.8%) already attended a review appointment after a previous mathematics exam where they could compare their received grades with the traditional grading schemes. All students said they attended to be better prepared the following time. However, 57.9% indicated they had problems understanding how grades were obtained using traditional grading.

To compare perceptions, we asked students to rank checkbox grading among other, more usual approaches to feedback (see Fig. 5) regarding comprehensibility. The results of this ranking exercise on two different exam tasks can be found in Table 5. On average, the traditional grading schemes were preferred above checkbox grading, classic written feedback and only a grade. Only a grade was by far the least preferred option.

The interview data does not fully confirm the perceived higher comprehensibility of traditional grading over checkbox grading during the ranking exercise in the questionnaire. The traffic light coding of traditional grading, displayed in Table 6, shows that only Sasha could independently link the traditional grading scheme with the grade she received in exam task 2. She answered $-5b \cdot (\cos(\alpha - 5) + i\sin(\alpha - 5)$ on 2(a) and correctly inferred she must have failed the entire task. When the researcher asked for explanations, she could immediately identify what should have been included in her answer. Two students could not draw correct conclusions when comparing their solutions against the traditional grading scheme. Jana, who submitted the wrong answer $-5[b \cdot (\cos\alpha + i\sin\alpha)] = -5b(\cos\alpha - 5 + i\sin\alpha - 5)$ and received 0/1.5, only noticed a superficial difference between her solution and the solution key but could neither link her mistake to the grading scheme nor suggest a grade:

Table 6 Results of the traffic light analysis of the cognitive processing of traditional grading

Sasha	Jana	Tom	Emile
2	2	2	2

"I don't really know. Because, well, I wrote α's without +180°, so instead of α + 180°, but they (the grading scheme, ed.) don't tell anything about this." (Jana)

Tom, who wrongly answered $-5b(\cos(\alpha) + i\sin(\alpha))$ on 2(a) and received 0/1.5, also failed to interpret the traditional grading scheme. First, he guessed he scored 0.75/1.5 because he noticed he forgot to write $+\pi$ in the argument of z_1. When the researcher intervened and asked whether 0.75 was a possible outcome based on the grading scheme (it was not), he changed his answer to 0.5/1.5. When the researcher asked to clarify his reasoning, he answered:

"I did not get zero because I have written a part of the solution. So yes, I would still say I received 0.5/1.5." (Tom)

Therefore, Tom misinterpreted the first criterion, which gave a partial score of 0.5 for transforming -5 to the correct polar form. Tom thought his grades were too low and believed he should have accrued some points anyway (*"I have written a part of the solution"*), as can also be seen in his interpretation of sub-task 2(b). Tom answered $b \cdot e^{ia} \cdot c \cdot e^{i3\beta} = bc(\cos(\alpha + 3\beta) + i\sin(\alpha + 3\beta))$, thereby using the Euler form of complex numbers, which is not part of the Flemish mathematics curriculum. The sub-task was graded 0.5/1. Asked to give a grade using the traditional grading scheme, Tom initially said he had full marks on the sub-task. When the researcher suggested this was not the case, he came to the correct conclusion that he forgot a third power in his modulus:

Table 5 Results of the ranking exercise on the comprehensibility of feedback types

Feedback type	Avg. rank	1st choice (%)	2nd choice (%)	3rd choice (%)	4th choice (%)
Traditional grading	1.58	58.2	27.0	13.5	1.3
Checkbox grading	1.90	36.5	43.2	14.9	5.4
Written feedback	2.67	5.3	29.7	58.1	6.8
Only a grade	3.86	0	0	13.1	86.5

"It seems I forgot the third power. Nevertheless, I used a nicer method (Euler form, ed.), and I think that deserves a little bit more appreciation. (...) And for the first sub-task: I still don't get why I got such a low score. I thought I would receive at least half a point." (Tom)

Finally, Emile, who wrongly answered $-5 \cdot b(\cos\alpha + i\sin\alpha)$ on sub-task 2(a) and was marked 0/1.5, immediately noticed he forgot a part in the argument of the polar form. So, he inferred he would receive 0.5/1.5 from the first criterion. When the researcher asked how he treated the -5 in his solution, he corrected his answer and correctly concluded that the first criterion did not apply and that he got 0/1.5.

5 Conclusions and discussion

In this study, we investigated the cognitive and affective processing of checkbox grading feedback (RQ1) and compared the perceptions of checkbox grading feedback with classical feedback approaches (RQ2). In order to do so, we distributed a questionnaire among all the 60 students who had participated in the mathematics exam at the Flemish Exam Commission, providing them access to sections of their checkbox grading feedback reports. Out of the group of 60 students, 36 took the time to complete the questionnaire, and 4 of them agreed to semi-structured interviews. These sample sizes are a limitation of the study, especially concerning the interviews, as no saturation can be expected with 4 participants (Hennink & Kaiser, 2022).

Concerning the first research question, the quiz and self-reports administered during the questionnaire and subsequent interviews were utilised. In regard to self-reports, it is noteworthy that a substantial proportion of students were positive with respect to checkbox grading: 97% of the students indicated the Flemish Exam Commission should adopt the approach, 91% of students concurred that the received feedback clearly indicates how the score was determined, 79% demonstrated an understanding of the relationship between the feedback and the score (which may also encompass the design choices made by the Flemish Exam Commission), 74% affirmed that they understand most of their feedback, and 77% indicated they will make even better exams based on their feedback. It is important to acknowledge that self-reports, while illuminating, may not fully capture the intricate cognitive processes that students engage in. Nonetheless, these findings suggest that the sampled students, at a minimum, perceive a degree of understanding and effectiveness in their interaction with checkbox grading feedback.

The quiz results, with an average score of 72%, exceeded expectations, particularly in contrast to the exam scores (*M*: 37%). These results were surprising since the quiz questions were intentionally designed to be challenging, necessitating a deep cognitive engagement with three complex components: the checkbox grading items, the examination of a fellow student's solution, and responding to quiz questions presented in the form of yes/no statements, all of which entailed substantial mathematical content. Furthermore, an interesting observation is the absence of a significant correlation between quiz performance and exam scores. This underscores the possibility that, despite the relatively low exam scores on average, a substantial proportion of the sampled students, including those with lower performance levels, could effectively decipher and interpret the checkbox grading feedback.

A consistent pattern emerged when connecting the outcomes of the quiz to the findings from the interviews. The four students could independently interpret their checkbox grading feedback for 11 of the 16 exam tasks discussed, constituting a success rate of 69%. Three distinct types of interpretation difficulties were identified for the remaining five exam tasks where an independent and accurate interpretation was not achieved. First, some students exhibited affective responses to the checkbox grading feedback, which obstructed a correct interpretation (e.g., Sasha on task 1, Tom on tasks 7 and 8). Such reactions are well-documented in the literature (Goetz et al., 2018; Koenka et al., 2021; Lipnevich & Smith, 2022), suggesting that the difficulties in understanding the feedback due to an affective response may not be inherently related to checkbox grading, and would possibly emerge in other feedback approaches too. The second interpretation challenge was insufficient content knowledge to comprehensively interpret the feedback (e.g., Tom on task 9, Emile on task 4). Notably, this was only a prevalent issue in two instances, considering that three out of the four interviewed students did not pass their respective exams. The final interpretation challenge emerged when a student disagreed with the grading criteria (Tom, task 8).

For the second research question, about the perceptions of checkbox grading feedback compared to classic approaches, a ranking exercise of equal feedback content, but formulated in different ways, was used in the questionnaire. Moreover, exam task 2 was discussed during the interviews using the traditional grading to compare the cognitive processing between traditional and checkbox grading.

Regarding the ranking exercise, the traditional grading schemes of the Flemish Exam Commission were predominantly favoured over other options, such as checkbox grading, classic written feedback, or only communicating a grade. Unsurprisingly, the latter option, simply communicating a grade, was the least preferred.

However, the position of classic written feedback is intriguing. These feedback comments were characterised by concise, personalised explanations of students' errors and omissions. Students only needed to invest effort in understanding these comments, which were straightforward and self-explanatory, without requiring them to relate their solutions to grading guidelines or solution keys. It is worth noting that this approach was less favoured compared to checkbox grading and traditional grading, both of which involve the application of grading criteria. This observation aligns with the findings of Harks et al. (2014), who concluded that "Learners receiving process-oriented feedback (written feedback) for the first time might struggle to deduce evaluation criteria from the unfamiliar, copious feedback message, to memorise them, or to apply them while evaluating their learning processes and outcomes" (p. 283). Another potential explanation for this preference is students' limited familiarity with receiving classic written feedback in the context of mathematics. Studies such as Knight (2003) have demonstrated that students in mathematics classes typically receive grades for their tests, with infrequent instances of mistakes being explicitly pointed out. Explanations accompanying these grades are seldom provided. It is also possible that students failed to discern a clear link between the grades assigned to the intermediate steps in classic written feedback, while traditional and checkbox grading distinctly illustrates how a grade is calculated.

A similar phenomenon may have influenced the preference for traditional grading over checkbox grading when comparing these two approaches. Students encountered checkbox grading reports for the first time during the ranking exercise, which might have led them to opt for what they were more accustomed to. Notably, the rank for traditional grading in first place somewhat contradicts the fact that 57.9% of students who had previously participated in a review appointment (involving traditional grading schemes) indicated that they required assistance in connecting their grades to the grading criteria. Simultaneously, 91% of students agreed that checkbox grading clearly indicated how a score is derived.

During the interviews, only one student (Sasha) could independently deduce the grade she received when comparing her solution to the traditional grading scheme, which depicts a more negative picture than the interview data on checkbox grading. Remarkably, all four students interviewed expressed their preference for checkbox grading without any prompting from the interviewer.

While our study was framed in a high-stakes mathematics exam at the Flemish Exam Commission, it is important to acknowledge the context-specific limitations of this research. Further research in larger-scale exams or real classroom settings with peer and self feedback could provide valuable insights into the value of checkbox grading in other contexts. Additionally, taking the behavioural processing ('What do students do with the received feedback?', see Sect. 2.2) and the resulting outcomes into account during a follow-up study would provide valuable insights into the generalisability of our findings, as this was not investigated.

In conclusion, this study explored students' perspectives on checkbox grading. It is essential to exercise caution when drawing conclusions due to the sample size. Nonetheless, the findings indicate that students generally view checkbox grading positively. They desire the approach to be implemented and demonstrate a notable comprehension of the feedback provided through checkbox grading. Notably, this positive perception seems consistent among students, as no significant correlation was observed with their respective exam scores. When examining preferences for different feedback approaches, students exhibit an intuitive preference for feedback presented in grading guidelines, as opposed to written feedback or mere grade communication. It is worth noting that students tend to favour traditional grading guidelines over checkbox grading. However, some limited evidence suggests that the traditional grading schemes may be more challenging to understand.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11858-024-01550-6.

Funding This research is funded by a doctoral fellowship (1S95920N) granted to Filip Moons by FWO, the Research Foundation of Flanders (Belgium).

Declarations

Conflict of interest No conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

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

Atkinson, D., & Lim, S. L. (2013). Improving assessment processes in Higher Education: Student and teacher perceptions of the

effectiveness of a rubric embedded in a LMS. *Australasian Journal of Educational Technology*. https://doi.org/10.14742/ajet.526

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

Bloxham, S., Den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: Exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education, 41*(3), 466–481. https://doi.org/10.1080/02602938.2015.1024607

Bloxham, S., & West, A. (2007). Learning to write in higher education: Students' perceptions of an intervention in developing understanding of assessment criteria. *Teaching in Higher Education, 12*(1), 77–89. https://doi.org/10.1080/13562510601102180

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

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 U. T. Jankvist, M. van den Heuvel-Panhuizen, & M. Veldhuis (Eds.), *Eleventh Congress of the European Society for Research in Mathematics Education* (Vol. TWG21, Issue 3). Freudenthal Group. https://hal.science/hal-02430515

Brunner, M., & Süb, H.-M. (2005). Analyzing the reliability of multidimensional measures: An example from intelligence research. *Educational and Psychological Measurement, 65*(2), 227–240. https://doi.org/10.1177/0013164404268669

Cartney, P. (2010). Exploring the use of peer assessment as a vehicle for closing the gap between feedback given and feedback used. *Assessment & Evaluation in Higher Education, 35*(5), 551–564. https://doi.org/10.1080/02602931003632381

Case, S. (2007). Reconfiguring and realigning the assessment feedback processes for an undergraduate criminology degree. *Assessment & Evaluation in Higher Education, 32*(3), 285–299. https://doi.org/10.1080/02602930600896548

Darabi Bazvand, A., & Rasooli, A. (2022). Students' experiences of fairness in summative assessment: A study in a higher education context. *Studies in Educational Evaluation, 72*, 101118. https://doi.org/10.1016/j.stueduc.2021.101118

Gamage, S. H. P. W., Ayres, J. R., & Behrend, M. B. (2022). A systematic review on trends in using Moodle for teaching and learning. *International Journal of STEM Education, 9*(1), 9. https://doi.org/10.1186/s40594-021-00323-x

Gawande, A. (2010). *The checklist manifesto: How to get things right* (1st ed.). Picador.

Gillham, B. (2005). *Research interviewing: The range of techniques*. McGraw-Hill Education.

Goetz, T., Lipnevich, A. A., Krannich, M., & Gogol, K. (2018). Performance feedback and emotions. In A. A. Lipnevich & J. K. Smith (Eds.), *The cambridge handbook of instructional feedback* (1st ed., pp. 554–574). Cambridge University Press. https://doi.org/10.1017/9781316832134.027

Harks, B., Rakoczy, K., Hattie, J., Besser, M., & Klieme, E. (2014). The effects of feedback on achievement, interest and self-evaluation: The role of feedback's perceived usefulness. *Educational Psychology, 34*(3), 269–290. https://doi.org/10.1080/01443410.2013.785384

Hennink, M., & Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social Science & Medicine, 292*, 114523. https://doi.org/10.1016/j.socscimed.2021.114523

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

Jonsson, A. (2013). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education, 14*(1), 63–76. https://doi.org/10.1177/1469787412467125

Jonsson, A., & Panadero, E. (2018). Facilitating students' active engagement with feedback. In A. A. Lipnevich & J. K. Smith (Eds.), *The cambridge handbook of instructional feedback* (pp. 531–553). Cambridge University Press. https://doi.org/10.1017/9781316832134.026

Kloosterman, P., & Warren, T. L. J. (2014). Can technology help in mathematical assessments? A review of computer aided assessment of mathematics. *Journal for Research in Mathematics Education, 45*(4), 534–537. https://doi.org/10.5951/jresematheduc.45.4.0534

Knight, N. (2003). Teacher feedback to students in numeracy lessons: Are students getting good value. *Set Research Information for Teachers, 3*, 40–45. https://doi.org/10.18296/set.0704

Koenka, A. C., Linnenbrink-Garcia, L., Moshontz, H., Atkinson, K. M., Sanchez, C. E., & Cooper, H. (2019). A meta-analysis on the impact of grades and comments on academic motivation and achievement: A case for written feedback. *Educational Psychology, 41*(7), 922–947. https://doi.org/10.1080/01443410.2019.1659939

Koenka, A. C., Linnenbrink-Garcia, L., Moshontz, H., Atkinson, K. M., Sanchez, C. E., & Cooper, H. (2021). A meta-analysis on the impact of grades and comments on academic motivation and achievement: A case for written feedback. *Educational Psychology, 41*(7), 922–947. https://doi.org/10.1080/01443410.2019.1659939

Lemmo, A. (2021). A tool for comparing mathematics tasks from paper-based and digital environments. *International Journal of Science and Mathematics Education, 19*(8), 1655–1675. https://doi.org/10.1007/s10763-020-10119-0

Lipnevich, A. A., Berg, D. A. G., & Smith, J. K. (2016). Toward a model of student response to feedback. In G. Brown & L. Harris (Eds.), *Handbook of human and social conditions in assessment.* Routledge.

Lipnevich, A. A., & Smith, J. K. (2022). Student—feedback interaction model: Revised. *Studies in Educational Evaluation, 75*, 101208. https://doi.org/10.1016/j.stueduc.2022.101208

Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability*. National Assessment Agency (UK). https://filestore.aqa.org.uk/content/research/CERP_RP_MM_01052005.pdf

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 and Education Open, 3*, 100086. https://doi.org/10.1016/j.caeo.2022.100086

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

Moons, F., Vandervieren, E., & Colpaert, J. (2023b). *Checkbox grading of handwritten mathematics exams with multiple assessors: field study on time, inter-rater reliability, usage & views*. Manuscript submitted for publication.

Moons, F., Holvoet, A., Klingbeil, K., & Vandervieren, E. (2024). Comparing reusable, atomic feedback with classic feedback on a linear equations task using text mining and qualitative techniques. *British Journal of Educational Technology. Advance online publication.*. https://doi.org/10.1111/bjet.13447

Morgan, C., & Watson, A. (2002). The Interpretative nature of teachers' assessment of students' mathematics: Issues for equity. *Journal for Research in Mathematics Education, 33*(2), 78–110. https://doi.org/10.2307/749645

O'Donovan, B., Price, M., & Rust, C. (2004). Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teaching in Higher Education, 9*(3), 325–335. https://doi.org/10.1080/1356251042000216642

Orsmond, P., Merry, S., & Reiling, K. (2002). The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education, 27*(4), 309–323. https://doi.org/10.1080/0260293022000001337

Panadero, E., Lipnevich, A., & Broadbent, J. (2019). Turning self-assessment into self-feedback. In M. Henderson, R. Ajjawi, D. Boud, & E. Molloy (Eds.), *The impact of feedback in higher education* (pp. 147–163). Springer International Publishing. https://doi.org/10.1007/978-3-030-25112-3_9

Price, M., Rust, C., O'Donovan, B., Handley, K., & Bryant, R. (2012). *Assessment Literacy: The Foundation for Improving Student Learning*. Oxford Centre for Staff and Learning Development.

Rust, C., Price, M., & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment & Evaluation in Higher Education, 28*(2), 147–164. https://doi.org/10.1080/02602930301671

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

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

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

Yan, T., Conrad, F. G., Tourangeau, R., & Couper, M. P. (2011). Should I stay or should i go: the effects of progress feedback, promised task duration, and length of questionnaire on completing web surveys. *International Journal of Public Opinion Research, 23*(2), 131–147. https://doi.org/10.1093/ijpor/edq046

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.