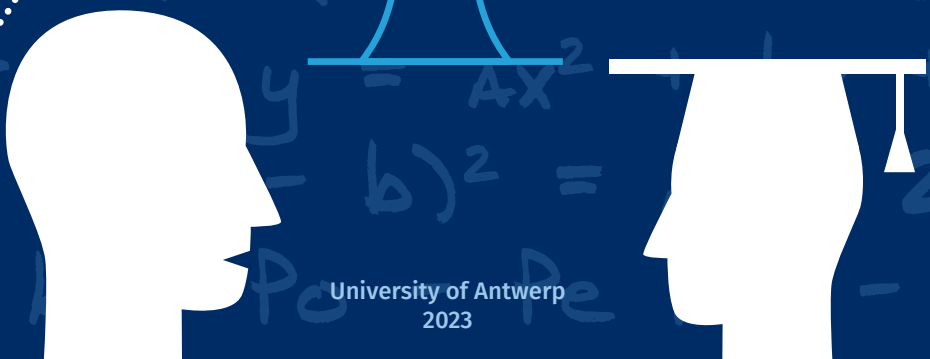
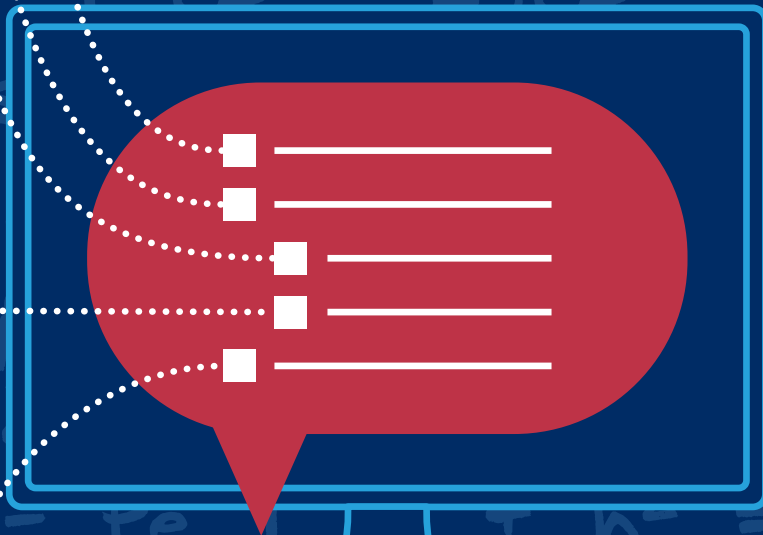


Semi-automated assessment of handwritten mathematics tasks

Atomic, reusable feedback
for assessing student tests by teachers
and exams by a group of assessors

Filip Moons



This work is dedicated to Roger Van Nieuwenhuyze, a remarkable mathematics teacher trainer, mentor & example, who passed away in the last month of this PhD. I will forever be grateful for your response when I told you in June 2010: 'I think I should leave', and you said, 'I think so too.'

© 2023 Filip Moons. All rights reserved.

Cover and graphic design: Caro Van Borm

Printed by: Pixartprinting

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

ISBN 978-9-05728-797-8
D/2023/12.293/21

Semi-automated assessment of handwritten mathematics tasks

Atomic, reusable feedback
for assessing student tests by teachers
and exams by a group of assessors

Filip Moons

Dissertation submitted for the degree of doctor in Educational Sciences at the University of Antwerp.

Supervisors: Prof. dr. Ellen Vandervieren & Em. prof. dr. Jozef Colpaert

Antwerp School of Education | Faculty of Social Sciences

Antwerp, 2023



Composition of the doctoral jury

Supervisors

Prof. dr. Ellen Vandervieren	University of Antwerp, Belgium
Em. prof. dr. Jozef Colpaert	University of Antwerp, Belgium

Doctoral committee

Prof. dr. Johan Deprez	KU Leuven, Belgium
Prof. dr. David Eelbode	University of Antwerp, Belgium

Doctoral jury

Prof. dr. Bärbel Barzel	University of Duisburg-Essen, Germany
Prof. dr. Christopher J. Sangwin	University of Edinburgh, United Kingdom

Chair

Prof. dr. Sven De Maeyer	University of Antwerp, Belgium
--------------------------	--------------------------------

FOREWORD

On a summer day in July 2018, I was lying in a hammock next to the Soča River (Slovenia), daydreaming about the future. I had been a mathematics teacher at the 'Hoofdstedelijk Atheneum Karel Buls' and an assistant at the teacher training of the University of Antwerp for a couple of years by then, but still, I felt there was so much left to learn. As a teacher trainer, I came across many exciting mathematics education studies. Still, I had no idea how the methodology behind them worked, nor had I had enough time to read them properly. The two professors I was an assistant for at the time, Prof. dr. Ellen Vandervieren and Prof. dr. Jozef Colpaert, talked me into doing a PhD a couple of weeks before. Jozef's advice kept popping into my mind: 'You're almost 30 now. You should do it now if you ever want to do a PhD. Every additional school year will make it harder to detach from practice and move into theory. Moreover, doing it now might also benefit you in the future.' Leaving the hammock, I decided: let's do this! Let us write a PhD project proposal next school year and see where it brings us.

Writing a PhD proposal and getting inspiration to know the spots of unexplored territory as a day-to-day mathematics teacher was more challenging than it seemed. However, a regular routine in my mathematics lessons was fascinating me. I often used two-stage tasks for challenging topics (e.g., proving geometric identities, convergence proofs of sequences or analysing functions). My experience with regular graded homework was not so positive. In times of Whatsapp, the solutions were shared within the class very quickly, with me ending up assessing students who simply copied each other. In two-stage tasks, the dynamic was completely different. In the first stage, students handed in their homework by taking pictures and uploading them on the e-learning platform of the school. Next, I gave process-oriented written feedback to them. In the second stage, they had to use my feedback to improve their solutions. Two things were repeatedly noted when implementing this activity. First, students dared to show their mistakes. While students in graded homework do everything in their might to show their best side, you could finally see where they were in the learning process. Second, although students did not work together, I noticed many mistakes occurred multiple times, allowing me to copy-paste already given feedback to newly assessed students. I was curious about this practice: was this 'reuse' of feedback researched? Could it save time for teachers, making the process of giving feedback less daunting? Did there exist tools for this? Surprisingly, my search yielded little.

Even more so, when spitting out the literature in preparation for the writing of my PhD proposal, I came across the book 'Computer-assisted Assessment in Mathematics'

of Sangwin (2013b). In its last chapter, ‘The future’, it mentioned semi-automated assessment of mathematics — assessment methods where teachers and computers work together to assess student’s work — as one of the open questions, as handwriting remains easier to express yourself in mathematics education than writing digitally; and computers probably could not take over the whole assessment process in mathematics education any time soon. A doctoral topic was born.

In the first months of the PhD, the first experiment had to be set up. Besides sketching out a good study design, developing a Moodle plugin, and sampling mathematics teachers, one thing became clear: some feedback is probably more reusable than others. The idea of *atomic feedback* emerged: a set of format requirements possibly helping teachers to write more reusable feedback. After all, reusability can only be judged after everything is written, making it not a very helpful concept during the writing process itself.

In 2021, the first experiment sparked the interest of the Flemish Exam Commission. Atomic feedback, could that be something they can give to their students taking mathematics exams too? As a researcher, I saw an opportunity to solve one of the open questions after the first experiment: is it possible not only to let teachers reuse their own feedback but also share feedback in groups of teachers? After some group thinking, we conceived an ambitious research plan to turn their traditional grading method for handwritten exams into a semi-automated one.

Looking back, it is this experiment which I’m most proud of. The semi-automated assessment approach lived up to the expectations of all parties involved; it led to the discovery of a new chance-corrected κ statistic to measure inter-rater reliability, which also served as a topic for my thesis for the master in Statistical Data Analysis, which I obtained at the University of Ghent as part of my doctoral programme.

The PhD thesis before you results from three and a half magnificent years of programming, conducting research, learning statistics, and meeting fellow researchers in the field of mathematics education. It has been a blend of everything I love doing, and I have had the incredible freedom to live, study, and work as I pleased. While I eagerly awaited its submission in the weeks leading up to the end, I now find myself reminiscing about this fairly carefree period with great wistfulness.

Enjoy your reading!

● **Filip Moons**
Antwerp
June 2023

ACKNOWLEDGMENTS

As it takes a village to raise a PhD student, I want to express my heartfelt gratitude to my fellow villagers. We start from the birth of this PhD project. My sincere appreciation goes to my supervisors Ellen and Jozef, who encouraged me to do a PhD in the first place. I will never forget the whole process of writing up the PhD proposal for FWO Strategic Basic Research. Near the deadline, the 1st of March 2019, my days were filled with teaching mathematics during the day and writing the proposal during the evening and early mornings. When I went to school, a new version was sent to Ellen for a new revision; I will always be grateful for this excellent cooperation. Also, the people from the 'Dive into a PhD'-course of the UA Antwerp Ann Aerts, Bruno Hoste and Frederik Verleysen are warmly thanked for their professional support in both the writing of the proposal and preparation for the oral defence. Without these people, you would not be reading the PhD in front of you. FWO is warmly thanked for believing in the project and granting a strategic basic research fellowship.

During my doctoral years, I encountered various people worth a big thank you. First, my supervisors, Ellen and Jozef: thank you for keeping me on track and being very complementary to each other. I thank my doctoral committee, Sven, Johan, and David, for their valuable insights and feedback throughout the process. Bärbel and Chris, thank you for being on my jury! Thanks to the students and my colleagues of the teacher training Gilberte, Carine and Johan. Also, thanks to Leen, Loan, Seppe and the two Drieses for being marvellous office mates (although I was not there that often).

Also, the research of this PhD was only possible with the contribution of several people. I am particularly grateful to Katie De Jonge (SGI Lennik) and her students for providing us with 60 tasks on linear equations and to the 45 mathematics teachers from all over Flanders, who spent an entire summer day at our university in 2020 assessing these tasks. Next, I would also like to thank Vladomeare Obolonsky for serving as a second rater in coding all of the feedback items as atomic/non-atomic for the first study as a student assistant. Also, a big shout out to the people of the Flemish Exam Commission, especially Dries Vrijssen, Griet Esprit and Carmen Streat, for the vibrant collaboration in developing the checkbox grading approach. Finally, I am grateful to Robin Roevens of the university's ICT department for setting up the server environment on which all research actions occurred.

When Covid-19 emerged, a lovely side project came from the Netherlands (Utrecht University) and Germany (University of Duisburg-Essen): Math@Distance. We surveyed mathematics teachers and their students in our three countries multiple times to see

.....

how distance teaching worked out and how it influenced teachers' attitudes and beliefs. Ellen and I participated as the 'Flemish' team', which was a splendid opportunity for me to get to know and learn from other math ed researchers. Paul, Bärbel, Daniel, Marcel, Heleen, Michiel and Layla: thank you for turning the pandemic into a learning opportunity!

Another opportunity was a direct result of the pandemic: as all face-to-face conferences were cancelled, I searched for a meaningful alternative. I found it in studying statics during my master's in Statistical Data Analysis at the University of Ghent from September 2020 till January 2023. Birgit, Dirk, Emiel, Helena, Leander, Lieselot, Nathan and Stijn, thank you for all the exciting collaborations on demanding statistical tasks and projects. Thanks to my advisor for the master's thesis, Prof. dr. Jan De Neve, from which [Chapter 3](#) of this PhD is a direct result.

Friends are the family we choose for ourselves. Axel, Bert, Johan, Katrin, Raf and Thijs, thank you for being six totally different people but wonderful best friends. Axel, thanks for the lovely remote working stays at your home in Girona (Catalonia, Spain), where we combined meeting up during the evening with productivity during the day. Bert, thank you for convincing me to pursue a PhD, for your help with the proposal and for our lovely meetings with your wife, Elke, my godchild Léo, Nino and Marleen. Johan, thanks for being a very loyal friend, for our relaxing leisure activities, for all the help with our new home and for being the great father of my godchild Fien. Katrin, as we share a passion for mathematics education research and Eurovision, I apologise for my thousands of Whatsapp messages asking for your opinion about my articles, research and the newly announced Eurovision candidates, and sorry for turning you into a Eurovision monster. Raf, our shared interest in campy music and politics always led to vibrant late-evening conversations over the phone and some wonderful trips; I think some lyrics of Abba best summarises our friendship: *"I ask in all honesty. What would life be? Without a song or a dance, what are we?"*. Thijs: although the time you are in Belgium is scarce, it is always fun when you come crashing into our place in Antwerp: without realising it; you are often an example to me.

My partner and family members can not be forgotten. Sacha, while I did my PhD, you finally became an oncologist after 13 intense years of studying; we made a lovely house in Antwerp and got engaged. Having such a partner on my side is terrific; you always ensure my feet are firmly grounded when I want to start flying. Lisette, as our kitty and lady of the house, you promoted yourself to this PhD's most loyal (and sleepy) supporter. Mummy, Dad, Annick, Wim and Elena: thank you for being the family that we are. Mummy, thank you for listening endlessly to my not-so-exciting PhD progress during remote working days.

The people from Uitwiskeling, VWWL (Flemish Association of mathematics teachers), Platform Wiskunde Vlaanderen (association of Flemish Mathematics) and the curriculum commission can also not be forgotten. They are a vital pillar in my connection to the field of mathematics education in Flanders. Ann, Didier, Els, Ellen, Gitta, Hannelore, Isabelle, Johan, Koen, Lotte, Filip and Mark: although the curriculum development was far from a smooth process, in hindsight, the drama was often an entertaining distraction from research work.

CONTENTS

General introduction	11
I Semi-automated assessment for individual teachers	25
1 Atomic, reusable feedback: a semi-automated solution for assessing handwritten tasks? A crossover experiment with mathematics teachers	27
1.1 Introduction	28
1.2 Atomic feedback	31
1.3 Materials & Methods	33
1.4 Results	40
1.5 Discussion	45
1.6 Further Research	48
2 Comparing reusable, atomic feedback with classic feedback on a linear equations task using text mining and qualitative techniques	51
2.1 Introduction	52
2.2 Materials & Methods	55
2.3 Results & Discussion	59
2.4 Conclusions	70
II Semi-automated assessment for a group of assessors	75
3 Measuring agreement among several raters classifying subjects into one-or-more (hierarchical) nominal categories. A generalisation of Fleiss' kappa	77
3.1 Introduction	78
3.2 Derivation of the proposed kappa statistic	85
3.3 Worked-out examples	93
3.4 Further research	101
3.5 Conclusion	102
4 Checkbox grading of handwritten mathematics exams with multiple assessors: study on time, inter-rater reliability, usage & views	105
4.1 Introduction	106

4.2	Research framework & questions	110
4.3	Methods & Materials	113
4.4	Results	118
4.5	Discussion	125
4.6	Conclusion	128
5	Checkbox grading of handwritten mathematics exams with multiple assessors: how do students react to the resulting atomic feedback? A mixed-method study	131
5.1	Introduction	132
5.2	Research framework & questions	135
5.3	Methods	138
5.4	Results	146
5.5	Discussion	151
5.6	Conclusion	153
	General conclusion	155
	Bibliography	167
	Summary (English)	179
	Samenvatting (Nederlands)	183
	Appendices	187
	Appendix A. Test on linear equations (including solution key)	187
	Appendix B. Pilot study design of the crossover experiment	189
	Appendix C. Teachers' survey Items based on TAM model (Chapter 1)	189
	Appendix D. Codebook of atomic feedback items	190
	Appendix E. Mathematics Exam of the Flemish Exam Commission	193
	Appendix F. Assessors' survey Items based on TAM (Chapter 4)	199

✓ GENERAL INTRODUCTION

Feedback and assessment are widely accepted as powerful engines of learning processes (Hattie & Timperley, 2007). However, feedback in mathematics classrooms is often limited to evaluative feedback, meaning that students get grades on tests and sometimes mistakes are highlighted, but with no word of explanation (Knight, 2003). Many policy reports from different countries and regions have pointed to this flaw. Reasons reported included large class sizes and time constraints (e.g., Enu, 2021; Gibson et al., 2015). Indeed, the Eurydice (2021) report pointed out that 49% of all the teachers in the European Union indicate having too much assessment work, which makes it the second biggest complaint about the teaching profession, after having too many administration tasks. The lack of feedback in mathematics education is also a returning observation in the yearly reports of the Flemish education inspectorate (2023):

“Feedback in mathematics classrooms is often too focused on the product with insufficient attention to the reasoning or arithmetical errors underlying the mistakes.” (Flemish education inspectorate, 2023, p.65, own translation)

Using digital assessment can be a solution to the need for more feedback. Indeed, technology has influenced various aspects of assessment in mathematics education in recent years, e.g. enabling sequences of automated test items that can easily be generated and assessed by a computer (Pelkola et al., 2018). However, a lot of automated assessments focus on lower-order goals, such as procedural fluency (Hoogland & Tout, 2018), instead of concentrating on higher-order thinking in mathematics, such as conceptual understanding, adaptive reasoning and strategic competence (Gravemeijer et al., 2017; Griffin & Care, 2015), which is essential for mathematical proficiency. Indeed, Kilpatrick et al. (2001) constructed a widely accepted model in which mathematical proficiency is viewed as a set of 5 intertwined strands: **conceptual understanding**, **procedural fluency**, **strategic competence**, **adaptive reasoning**, and **productive disposition** (see Figure 1). Successful mathematics learning can only be achieved by focusing on all five strands, and — under the slogan ‘What You Test Is What You Get’ (WYTIWYG) — assessment and feedback should appeal to all of these competencies (Burkhardt, 1985).

The limited focus of fully-automated assessment on higher-order thinking goals stems from the fact that digital task environments offer too few mathematical tools that allow students to express themselves mathematically, as they would with paper-and-pencil (Drijvers, 2018). Most digital questions limit students to answer in pre-defined response

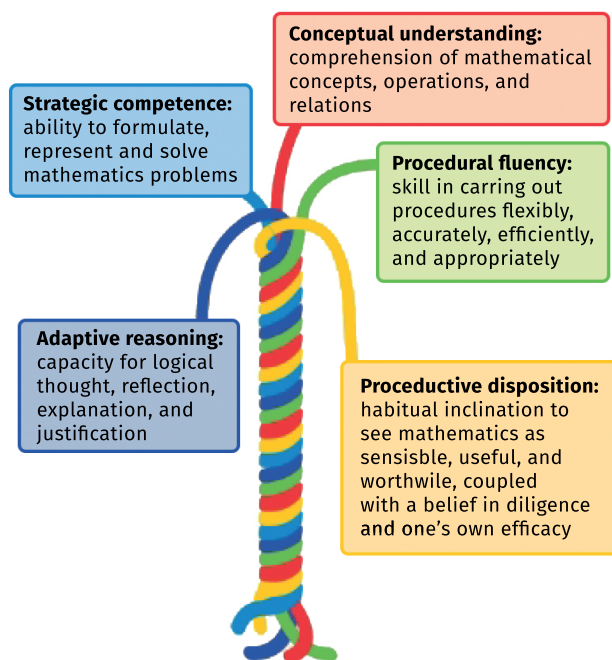


Figure 1 – The 5 intertwined strands of mathematical proficiency (Source: Kilpatrick et al., 2001; Picture: van de Walle et al., 2018)

fields (Bokhove and Drijvers, 2010; see Figure 3). In contrast, paper-and-pencil-based assessment (PP) allows students to express themselves more freely. Moreover, teachers' adoption of fully-automated assessment is linked to their beliefs, with some teachers fearing the loss of by-hand skills when implementing it into their classrooms (Thurm & Barzel, 2021).

This doctoral thesis wants to bridge the gap between paper-and-pencil-based assessment (PP) and fully-automated assessment (FA) by researching the possibilities of semi-automated assessment (SA), see Figure 2.

ABBREVIATIONS • In the rest of the introduction, we use the abbreviation **SA** to refer to semi-automated assessment. Fully automated assessment is indicated with **FA**, and **PP** refers to paper-and-pencil-based assessment.

Semi-automated assessment wants to combine the strengths of PP and FA while avoiding their weaknesses. First, SA leaves the idea that everything in computer-aided assessment has to be assessed fully automated and reinstates human intervention. This re-opens the possibility of letting students solve higher-order thinking questions with paper and pen. However, the difference with PP assessments lies in how assignments are assessed: instead of assessing those handwritten solutions manually, they are assessed digitally, taking advantage of the observation that students often make identical or analogous mistakes. There is much evidence that students do so: systematic error patterns in answers to math questions are already investigated thoroughly in literature

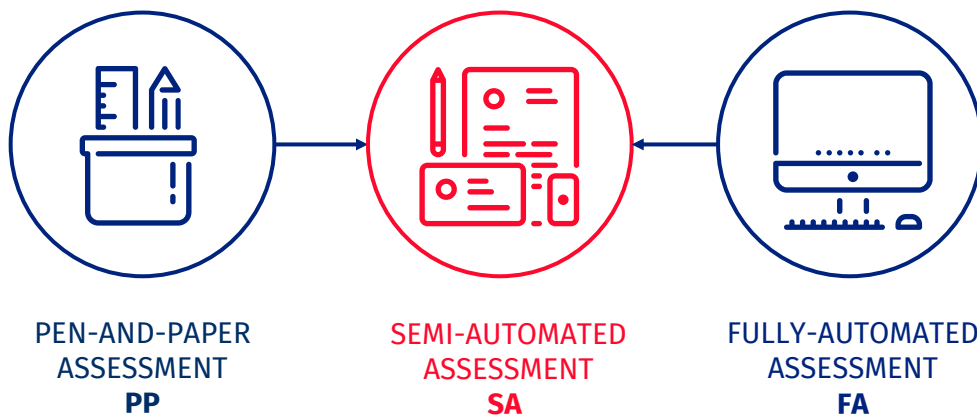


Figure 2 – Bridging the gap

(e.g. Movshovitz-Hadar et al., 1987; Schnepfer and McCoy, 2014). However, using these error patterns to speed up the assessment process remains surprisingly unstudied. When students make similar mistakes, we often suspect teachers to repeat themselves when giving specific step-level feedback (VanLehn, 2006) on the used solving method of a particular question. SA makes it possible to create a system that remembers given feedback on a specific question, making it very easy to reuse already given feedback. When the same mistake pops up more than once, the same grading and feedback can be reused quickly. The need for research on the topic was already mentioned by Sangwin (2013a) in the concluding chapter ‘The future’ in his book ‘Computer-Aided Assessment of Mathematics’. In 2022, Kinnear et al. reaffirmed this research gap in a collaboratively derived research agenda for e-assessment in undergraduate mathematics. One of the themes of this research agenda is ‘free-form student input’ with research question 40: *How can the suitability of e-assessment tools for summative assessment be improved by combining computer-marking and pen-marking?* directly relating to this work. Also, research questions 42 (*‘How can we automate the assessment of work traditionally done using paper and pen?’*) and 43 (*‘How can we emulate human marking of students’ working, such as follow-on marking and partially correct marking?’*) will be partly answered in this dissertation.

In this dissertation *feedback* and *assessment* are frequently used words. It is, therefore, appropriate to define them. A definition of *feedback* that fits the research conducted in this PhD is the one of Hattie and Timperley (2007): “Feedback is information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one’s performance or understanding” (p. 102). *Assessment* is defined as both the product and the process in which a teacher (assessor) judges a student’s task, which is often called a *diagnostic judgement* in the literature (Artelt & Rausch, 2014; Loibl et al., 2020). Feedback and assessment are sometimes intermingled in this dissertation. Still, a rule of thumb was used for this: if we want to stress the process of judging/grading a student’s task, we use *assessment*; if we want to stress the process of drafting a feedback message, we use *feedback*. In almost all circumstances, feedback is used if we want to make clear we go beyond just communicating a score. The assessment and feedback can be either summative or formative (Benett, 2011), depending on the chapter or how the SA approach is deployed in the classroom by a teacher.

In the following paragraphs, we first contrast SA with FA, present similar approaches to our proposed SA method, introduce the overarching research goals and conclude with an outline of the dissertation.

SEMI-AUTOMATED ASSESSMENT VS FULLY-AUTOMATED ASSESSMENT

The search for fully-automated assessment has been an established research field for some time (Bugbee, 1996; Denton et al., 2008), within the context of various subjects. The benefits of FA include extensive, immediate feedback for students and substantial time savings for teachers. It also offers nearly endless training possibilities in a variety of areas (e.g., science, mathematics), as questions can often be generated automatically (Sangwin, 2013b).

Despite its potential benefits, FA feedback is not appropriate for assessing all tasks. When simply transforming traditional paper-based tasks (PP) into digital formats, the so-called *migratory approach* (Ripley, 2009), the primary constraint in mathematics is that most FA systems can evaluate only the final answer. Because these answers are often expressed through multiple-choice inputs or inputs with pre-defined response fields (Bokhove and Drijvers, 2010; see Figure 3), it can be difficult or impossible for students to show their thinking process. As a result, there is no evidence to determine whether students have used the appropriate solving methods (Sangwin & Köcher, 2016). Moreover, multiple-choice questions can often lead students to solve questions in the inverse direction—determining whether the response options are appropriate instead of seeking the correct answer (Sangwin & Jones, 2017). Threlfall et al. (2007) observed that some questions are more natural for students to solve using paper and pencil. Lemmo (2021) confirms that there is a substantial difference in terms of assessment and

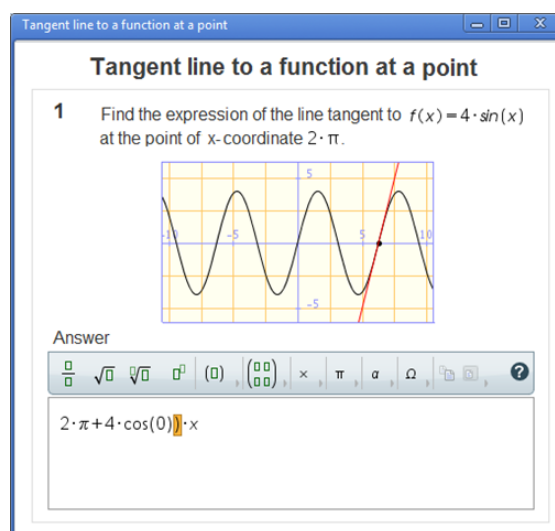


Figure 3 – Example FA question with pre-defined response field

thinking processes when a PP question is migrated into an FA question and provides a tool to identify the differences.

Applying SA techniques might offer a suitable compromise between FA and PP by resolving the current limitations of both PP and FA for some question types. As such, it challenges the existing frameworks of technology-rich assessment in mathematics education (Fahlgren et al., 2021). Firstly, SA could potentially offer significant time savings compared to PP, as it maximizes the reuse of previously entered feedback when assessing students' solutions. Secondly, SA allows students to write down any mathematical expression on a sheet of paper, using the structures they prefer for their reasoning, thereby showing all of their solution steps. As such, SA does not limit the use of open-ended, challenging questions involving higher-order thinking, as there are no pre-defined response fields. Thirdly, because the teacher retains control of the correction work and can write feedback whenever a mistake occurs, the need to develop complex correction schemes to anticipate mistakes in advance (as is the case for FA; Sangwin, 2013b) is eliminated. Although the loss of immediate feedback is a drawback, no significant differences have been found between delayed and immediate feedback for questions involving higher-order thinking (Van der Kleij et al., 2015). In many cases, SA might thus offer a valid assessment method that combines the advantages of PP and FA assessment (see Table 1)

Feedback & Assessment		
Paper-based (PP)	Computer-assisted assessment	
	Semi-automated (SA)	Fully automated (FA)
- delayed feedback	- delayed feedback	+ immediate feedback
+ natural mathematical expressions	+ natural mathematical expression	- mathematical expression difficult or impossible
+ no pre-defined response fields	+ no pre-defined response fields	- pre-defined response fields
+ easy question development	+ easy question development	- need for an automated correction scheme and anticipation of mistakes
+ higher-order thinking questions possible	+ higher-order thinking questions possible	- higher-order thinking questions difficult
- feedback possibly inconsistent between students	+ consistent feedback across students	+ consistent feedback across students
- time-consuming	+ time-saving	+ time-saving

Table 1 – Envisioned advantages and disadvantages of various types of assessment

SIMILAR APPROACHES

In recent years, several approaches have been developed that share the same goal as our SA approach: increasing efficiency in correcting handwritten tasks. In this section,

we provide a brief overview of these developments.

One well-known approach is the use of *rubrics*: assessment schemes that employ fixed sets of criteria (Moskal, 2000). Teachers indicate the relevant criteria for student tasks so that students can see which criteria need improvement. Rubrics thus provide feedback to the learner. They have nevertheless been the subject of criticism. For example, Sadler (2009) argues that most rubrics do not adequately represent the full complexity of qualitative judgements based on multiple criteria. Concerning problem-solving skills in science education, Docktor and Heller (2009) observe tension and trade-offs between global rubrics that can be applied to various problems and problem-specific rubrics that also consider the specific reasoning methods that are needed to solve a given problem. Our proposed SA method resolves this tension by saving feedback items so that teachers can select only relevant items to a particular student and add additional items if necessary to provide more detail. Unlike the one-size-fits-all feedback texts produced by rubric-based criteria, the SA method offers particular items for particular details, thus ensuring that the feedback is sufficiently specific to the student's responses to the task.

Another solution to the tension observed by Docktor and Heller (2009) has been proposed by Hull et al. (2013), who suggests starting with a global rubric and adding criteria specific to the problem at hand. A more flexible but similar idea is found in the *dynamic rubrics* used in the software tool Gradescope (Singh et al., 2017; see Figure 4). A dynamic rubric is composed of one or more rubric items, and more rubric items can be added throughout the correcting process. Each rubric item contains a grade and a description. The teacher's job is to select the items relevant to student solutions and, if necessary, add new rubric items for each new type of error. Such dynamic rubrics differ from traditional rubrics, as students do not encounter a grid of pre-set criteria but only a list of rubric items relevant to their solutions. Dynamic rubrics are similar to our SA approach in some respects, particularly when only a few (about 10) items are constantly reused. Nevertheless, Gradescope focuses primarily on grading handwritten tasks with a small set of criteria, with less attention to feedback. It thus becomes somewhat cumbersome when using a large number of rubric items (e.g., the screen becomes disorganised, and the grading process becomes tiring). For this reason, Gradescope is not appropriate for providing process-oriented formative feedback (Rakoczy et al., 2013). In contrast, our proposed SA system allows an almost infinite number of reusable feedback items to be presented to the student neatly in a hierarchical and personal list.

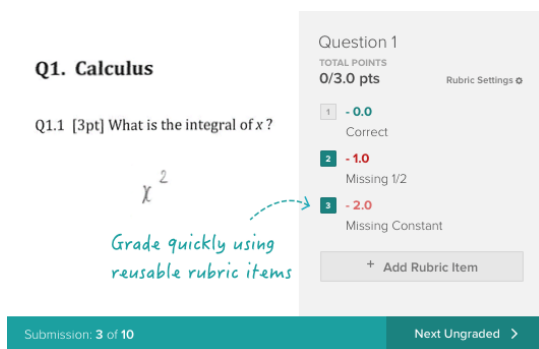


Figure 4 – Gradescope

Another similar approach consists of the ‘*frequently used comments*,’ seen in some tools. For example, the integrated Google Docs in Google Classroom (Google, 2021) features a so-called ‘statement bank’ (Denton & McIlroy, 2018). When giving feedback on assignments, a teacher can leave comments on student work and maintain a comment bank to store commonly recurring remarks (see Figure 5). The comment bank is nevertheless subject to the same drawback as Gradescope: it is manageable only with a limited number of items, and it lacks a sound suggestion system. This approach could thus also benefit from the ideas presented in this study.

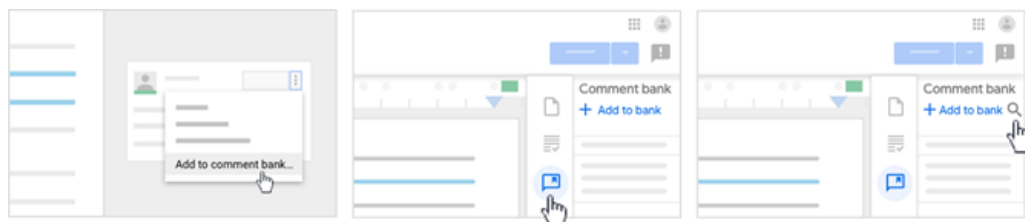


Figure 5 – Google Classroom

The first tools using artificial intelligence (AI) to analyse the handwritten solutions of students are now emerging. For example, Graide (Stanyon et al., 2022) facilitates optical character recognition (OCR), which converts a handwritten text into a digital solution (among other input methods). The internal AI system then applies ‘replay grading’ to determine whether the solution (or part thereof) has been seen before. If so, the system automatically grades the response or suggests feedback. If the solution is new, the teacher grades it, and the internal AI learns it. Although this approach is not yet flawless, it is promising.

RESEARCH GOALS

Building upon the existing literature and prior approaches discussed above, we invented two semi-automated assessment methods for mathematics education, corresponding to the two parts of this dissertation. The first method consists of the initial idea behind this PhD project: when teachers assess (scanned) handwritten solutions using a computer, the computer can save their feedback, so teachers can reuse it when following students who make similar mistakes. During a crossover experiment with 45 mathematics teachers, this approach was investigated from the teacher’s perspective and the feedback’s content. A first problem arose with the concept of reusing feedback: if a teacher writes feedback, is it reusable as such? An answer to that problem led to the idea of *atomic feedback*, which runs like a thread through the two parts and is introduced in the first chapter. The first study prompted many further research ideas and sparked the interest of the Flemish Exam Commission mathematics department. For years, they had been looking for ways to efficiently offer their students something more than just a grade on their exam. It led to a research collaboration and the second part of this PhD. This second semi-automated assessment method builds further on the first one by trying to answer the question: If teachers can reuse their own feedback, can a group of teachers share and reuse their feedback too? Given the specific context of the Flemish Exam Commission, we will speak of ‘assessors’ instead of ‘teachers’ in the second part. The second semi-automated assessment method will be called ‘checkbox

grading', which is suited to grade and give feedback to high-stakes mathematics exams. The approach is examined from the perspective of both assessors and students. From the assessors' perspective, we studied how the grading method influenced their time investments, inter-rater reliability, usage and views. From the student's perspective, we investigated whether they understood the resulting feedback. In both parts, paper-and-pencil (PP) assessment approaches serve as a reference to which the new methods were compared. [Chapter 3](#) stands out from the rest of the dissertation because it introduces a novel statistical measure for analysing inter-rater reliability. A research question in [Chapter 4](#) led to the search and eventually the discovery of this new measure, which is then used in [Chapter 4](#) to answer the research question.

Based on the initial targets stated in the application for an FWO strategic basic research fellowship and across both parts of this doctoral thesis, five overarching research goals guided the project:

- [RG 1] Is it possible to develop software to assess and give feedback to mathematics tasks semi-automatedly?
- [RG 2] Does a semi-automated approach with reusable feedback lead to time savings compared to paper-and-pencil feedback?
- [RG 3] Does a semi-automated assessment approach influence grading reliability?
- [RG 4] How does a semi-automated approach with reusable feedback influence the given feedback regarding characteristics, content and quality?
- [RG 5] How do the different users (teachers, assessors and students) perceive semi-automated assessment and feedback?

The connection between these overarching research goals and the research questions investigated in each chapter is depicted in [Figure 6](#), which serves as a graphical outline of this dissertation. The research goals and the graphical outline will also structure our general conclusion at the end of the dissertation. Every research question is directly linked to a research goal. The reverse is invalid: there is no connection between research goal 1 ([RG 1]) and a research question: although the software development was foundational for both parts, it was never a direct research object. Given that it does represent a substantial but somewhat invisible part of this PhD, we will discuss it in the general conclusion. It is also a key element in the valorisation of the research results.

Additionally, [Figure 6](#) also gives an indirect overview of the blind spots that remain. While grading reliability ([RG 3]) is a central issue in the second part, it has not been researched for the first semi-automated approach. Additionally, the goal of feedback quality ([RG 4]) is investigated differently in both parts: in the first part, the given feedback was analysed using text-mining and qualitative techniques; in the second part, feedback quality was directly measured by its recipients: the students. All these blind spots serve as fruitful ideas for further research.

In the following section, we briefly outline the different chapters.

	Research question	Research goal
Part I: Semi-automated assessment for individual teachers		[RG 1]
1. Atomic, reusable feedback: a semi-automated solution for assessing handwritten tasks? A crossover experiment with mathematics teachers	[RQ 1.1] [RQ 1.2] [RQ 1.3] [RQ 1.4]	[RG 2] [RG 4] [RG 4] [RG 5]
2. Comparing re-usable, atomic feedback with classic feedback on a linear equations task using text mining and qualitative techniques	[RQ 2.1]	[RG 4]
Part II: Semi-automated assessment for a group of assessors		[RG 1]
3. Measuring agreement among several raters classifying subjects into one-or-more (hierarchical) nominal categories. A generalisation of Fleiss' kappa	[RQ 3.1]	[RG 3]
4. Checkbox grading of handwritten mathematics exams with multiple assessors: study on time, inter-rater reliability, usage & views	[RQ 4.1] [RQ 4.2] [RQ 4.3]	[RG 2] [RG 3] [RG 5]
5. Checkbox grading of handwritten mathematics exams with multiple assessors: how do students react to the resulting atomic feedback? A mixed-method study	[RQ 5.1] [RQ 5.2]	[RG 5] [RG 4]

Figure 6 – Outline of the dissertation and the link between the research question and overarching research goals

OUTLINE OF THE DISSERTATION

Part I: Semi-automated assessment for individual teachers

1. Atomic, reusable feedback: a semi-automated solution for assessing handwritten tasks? A crossover experiment with mathematics teachers

The first chapter introduces the concept of atomic feedback. Next, it describes the first crossover experiment with 45 mathematics teachers (9 in the pilot study, 36 in the actual study) with semi-automated assessment. They assessed 60 linear equations tasks from students in two conditions: SA and PP. In the SA condition, they used a self-developed Moodle plug-in that stored their feedback items to enable reusing them when following students make similar mistakes. In the PP condition, they only received a text box to give their feedback. In the SA condition, teachers were taught to write their feedback atomically. We investigated the differences between SA and PP regarding time investment and the amount of feedback. Moreover, we explored if we could distinguish atomic feedback from non-atomic feedback, if following the guidelines of atomic feedback makes feedback more reusable and, finally, how teachers used and perceived the developed SA tool.

2. Comparing reusable, atomic feedback with classic feedback on a linear equations task using text mining and qualitative techniques

Comparing and contrasting the feedback given in both conditions during the crossover experiment with 45 mathematics teachers is what is done in the second chapter. More generally, the chapter seeks to answer how the possibility of reusing feedback influences the feedback's characteristics regarding form, content and quality. Two methodological approaches were used: text mining and qualitative analysis. Text mining uses computer algorithms to identify meaningful patterns and insights in written text. The qualitative analysis used a codebook from the mathematics education literature for classifying feedback. It looked into the concreteness, the focus of the diagnosis, the diagnostic activity and quality features of the diagnosis of the given feedback in both conditions. By combining both methodological approaches, the papers also serve as an example of what text mining can('t) offer for educational research on feedback.

Part II: Semi-automated assessment for a group of assessors

3. Measuring agreement among several raters classifying subjects into one-or-more (hierarchical) nominal categories. A generalisation of Fleiss' kappa

The third chapter introduces a new chance-corrected κ statistic to measure inter-rater reliability. It is a direct result of a research question investigated in [Chapter 4](#), in which we wanted to compare the inter-rater reliability of blind versus visible checkbox grading ([[RQ 4.2](#)]). We needed an inter-rater reliability measure to answer the research question, allowing raters to classify subjects into one or more hierarchical categories. Well-known measures such as Cohen's kappa or Fleiss' kappa were not suited as they only allow a rater to classify a subject into exactly one category. Surprisingly, such a measure was lacking in the literature. This statistical and methodological chapter solves that gap by presenting a generalisation of Fleiss' kappa, allowing the selection of multiple

categories for subjects by raters. The proposed κ statistic is then used in the next chapter to answer the research question.

4. Checkbox grading of handwritten mathematics exams with multiple assessors: study on time, inter-rater reliability, usage & views

The fourth chapter is the first of two chapters focusing on checkbox grading. Checkbox grading is a newly devised semi-automated grading approach for high-stakes exams. Every assessor receives a list of checkboxes and needs to tick those that apply to the student's solution. Dependencies between these checkboxes can be set to ensure all assessors take the same path down the grading scheme. The system automatically calculates the grade and generates feedback for the student, giving a detailed insight into what went wrong and how the grade was obtained based on predefined atomic feedback. The grading approach was developed in close cooperation with the Flemish Exam Commission and is somewhat the 'group equivalent' of the SA tool in the first part, as the checkbox items are also written in an atomic way. The main difference between the two approaches is that the items are not created while assessing a sequence of students, but are predefined by, in this case, the exam designers of the Flemish Exam Commission. In this chapter, we look at the new semi-automated grading approach from the assessors' perspective: How does the approach influence their time investment, inter-rater reliability, and how did they perceive and use it? The answer to these research questions was found using a mainly quantitative approach. The context was a real exam organised and designed by Flemish Exam Commission on advanced mathematics.

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

The last chapter investigates the student's perspective on checkbox grading. When all their exams were assessed a month after taking the exam, we investigated students' cognitive and behavioural processing of the given checkbox grading feedback using a questionnaire and semi-structured interviews. In the questionnaire, we polled their view on current feedback practices at the Flemish Exam Commission; they had to rank four different feedback types from most to least comprehensible, had to take a quiz with true/false questions about their comprehension of the feedback given to a fellow student and at the end, they received their personal checkbox grading feedback on three exam tasks, and we polled their understanding and views on this feedback. Of the 60 students who took the exam, 36 participated in the questionnaire. Four agreed to semi-structured interviews in which we investigated more deeply the cognitive processes that took place when they tried to make sense of their received checkbox grading feedback using a think-aloud protocol. Moreover, we compared it to their processing of the traditional feedback usually given at the Flemish Exam Commission.

PhD by publication

All five chapters are based on research articles that have been published ([Chapter 1](#)), have been published as a preprint ([Chapter 3](#)) or are currently under review ([Chapters 2, 4 and 5](#)). The research article on which the chapter is based is always indicated on the chapter's title page in bold. Almost all chapters were also (partly) presented at

research conferences, and the resulting (peer-reviewed) proceeding papers are also mentioned. This 'PhD by publication'-characteristic implies some repetition and overlap of information but also ensures that the chapters are readable in isolation. Every chapter starts with highlights described in layman's terms (left on every title page). These highlights serve as a common thread throughout the dissertation, glueing all the papers together. On top of that, they make the research outcomes available for a broad audience or readers with little reading time. At the end of the dissertation, in the general conclusion, we look back at our overarching research goals declared above and critically analyse them in light of the outcomes of the different studies, and we will give directions for future research.





PART 1

**SEMI-AUTOMATED
ASSESSMENT FOR
INDIVIDUAL TEACHERS**

HIGHLIGHTS

- ✓ Crossover experiment investigating how the feedback process for handwritten tasks can be made more efficient by having teachers and computers collaborate.
- ✓ Semi-automated approach with reusable feedback: when a teacher writes feedback for a student, the computer saves it, so that it can be reused when subsequent students make the same or similar mistakes.
- ✓ Writing 'reusable' feedback by introduction of atomic feedback: a set of form requirements for feedback items, demonstrating that it makes feedback significantly more reusable.
- ✓ Remarkable result: the semi-automated approach led teachers to give significantly more feedback instead of saving time.



✓ Chapter 1

ATOMIC, REUSABLE FEEDBACK: A SEMI-AUTOMATED SOLUTION FOR ASSESSING HANDWRITTEN TASKS? A CROSSOVER EXPERIMENT WITH MATHEMATICS TEACHERS

PUBLICATIONS

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. (2021). Writing atomic, reusable feedback to assess handwritten math task semi-automatedly. In U.T. Jankvist, R. Elicer, A. Clark-Wilson, H.-G. Weigand, & M. Thomsen (Eds.), *Proceedings of the 15th International Conference on Technology in Mathematics Teaching (ICTMT15)*, 13-16 September 2021 in Copenhagen, Denmark, pp. 87–88. Aarhus University. <https://doi.org/10.7146/aul.452>

Moons, F., & Vandervieren, E. (2021). Atomic, reusable feedback: a technology-mediated solution for assessing handwritten math tasks? *Proceedings of the 14th International Congress on Mathematical Education (ICME-14)*, 11-18 July in Shanghai, China. <https://www.icme14.org/ueditor/jsp/upload/file/20210706/1625550198643058915.pdf>

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

ABSTRACT

Feedback has been recognised as a crucial element in the learning and teaching process. Although teachers know and accept this, they are not always eager to engage in this tedious and time-consuming activity. This study investigates how computers can work together with teachers to make the process of giving feedback more efficient by introducing a semi-automated approach (SA) with reusable feedback: when a teacher writes feedback for a student, the computer saves it, so it can be reused when following students make similar mistakes. We devised the concept of atomic feedback, a set of form requirements that could enhance feedback's reusability. To write atomic feedback, teachers have to identify the independent errors and write brief feedback items for each separate error. Our SA approach with reusable feedback was implemented in Moodle. During a crossover experiment with math teachers ($n = 36 + 9$ in pilot study), we examined (1) whether SA saves time or changes the amount of feedback, as compared to traditional, paper-based correction work, (2) the extent to which the feedback was atomic, (3) whether atomic feedback enhances the reusability of feedback and (4) how teachers used and perceived the SA system. In light of the results, which suggest that atomic feedback is indeed reusable, we propose formal requirements for writing reusable feedback. Nevertheless, teachers did not save time using the SA system, but they provided significantly more feedback.

1.1 INTRODUCTION

Feedback is being increasingly recognized as an essential part of the learning and teaching process (Hattie & Timperley, 2007; Wisniewski et al., 2019). As observed by Shute (2008), "The premise underlying most of the research conducted in this area is that good feedback can significantly improve learning processes and outcomes, if delivered correctly." (p. 154) If delivered correctly, feedback allows teachers to encourage and reward their students while challenging and supporting learners within a scaffolding environment.

Feedback is a multifarious concept. It may or may not be linked to formal evaluation in terms of grading and assessment. As an iterative process, it may require several cycles before final (summative) grading. It can be given either more or less formally, on an individual or collective basis, face-to-face or remotely, synchronously or asynchronously, in a terse or detailed manner, and either explicitly or implicitly.

Feedback is of significant motivational value to learners, as supported by a considerable body of evidence. Theoretical support can also be found in motivation theories and models, including self-determination theory (Deci & Ryan, 2000), the ARCS model (Keller, 2009), expectancy-value theory (Wigfield & Eccles, 2000), self-efficacy theory (Bandura, 1977), goal-setting theory (Locke & Latham, 2013), and attribution theory (Weiner, 1986).

Although teachers acknowledge the value of feedback, they often find the evaluation, correction, and grading of student work to be tedious and time-consuming, and they are not always eager to engage in such activities. While they teach with passion, teachers have little motivation to provide timely, detailed feedback. For example, 49% of the teachers in the European Union complain about having too many correcting/grading tasks (Eurydice, 2021). Similarly, as reported by Gibson et al. (2015), 53% of all British teachers complain about the workload associated with grading and providing feedback.

The repetitive nature of feedback is a main problem, as most student answers contain systematic error patterns, meaning that different students often make similar mistakes (Movshovitz-Hadar et al., 1987; Schnepfer & McCoy, 2014). In paper-based assessment processes (PP), teachers try to provide targeted, relevant, original, and personal feedback. Due to such systematic errors, however, they must often repeat the same comments multiple times, leading to one of the main ideas of this chapter: pieces of feedback can often be reused for multiple students. A system that facilitates this reuse would help eliminate such repetition, possibly allowing considerable time savings and improved feedback while enhancing student results.

Digital learning environments simplify administrative tasks for teachers by allowing students to access and execute tasks and exercises online. Teachers can edit, correct, and grade their work anytime and anywhere, and computers can grade many types of exercises, including half-open question types, automatically. The search for fully automated (FA) assessment has been an established research area for decades (Bugbee, 1996; Denton et al., 2008; Sangwin, 2013b). The growing popularity of FA assessment raises questions concerning whether such systems can provide valid assessments of the full spectrum of skills. Within the context of mathematics education, Hoogland and Tout (2018) warned that digital assessment tends to focus on lower-order goals. They therefore argue that PP methods are better suited to questions requiring higher-order thinking, as they allow teachers (unlike computers) to assess the entire thinking process. These methods also allow students to express themselves more freely (Bokhove & Drijvers, 2010).

ABBREVIATIONS • In the rest of this chapter, we use the abbreviation **SA** to refer to our semi-automated assessment approach with reusable feedback in which handwritten tasks are corrected using a computer. Fully automated approaches are indicated with **FA**, with **PP** referring to traditional, paper-and-pencil tasks, also corrected paper-based by teachers.

In this study, we sought to devise a method for bridging the gap between fully automated (FA) computer-based assessment and purely paper-and-pencil (PP) feedback, thereby developing a new, semi-automated (SA) feedback method. In the proposed semi-automated (SA) feedback system, students work out their solutions with paper and pen, but the teacher assesses them digitally. When a teacher writes feedback for a student, the computer saves it, so that it can be reused for subsequent students making the same or similar mistakes. The proper formulation of such *reusable* feedback nevertheless poses a challenge. When expressed as a traditionally written text consisting of interdependent phrases (Winstone et al., 2017), feedback tends to be highly targeted at specific students (Glover & Brown, 2006), which compromises its reusability.

Moreover, the criterion of reusability can be assessed only after all correction work has been completed, and it provides no guidance to teachers as they are writing the feedback. To address this problem, we devised *atomic feedback*: a collection of form requirements intended to make written feedback more reusable (see [section 1.2](#)). The SA system, which suggests relevant items for reuse, was developed in Moodle (Gamage et al., 2022).

1.1.1 Research questions

This chapter makes a first attempt to investigate our proposed SA idea. Designed as an experiment, this crossover study was conducted among 36 mathematics teachers (+ 9 teachers in the pilot study) giving feedback on 60 tasks in two conditions: SA and PP. In the SA condition, teachers used the SA tool, which allowed feedback to be expressed atomically, with each item being saved and made available for reuse with subsequent students. In the PP condition, teachers were presented with only a textbox in which to express feedback as if they would have on a sheet of paper. The study investigates four research questions:

[RQ 1.1] What is the difference between SA and PP with regard to time investment and amount of feedback?

[RQ 1.2] Is it possible to create content-specific guidelines for atomic feedback to distinguish atomic feedback items from non-atomic items in the SA condition?

[RQ 1.3] How reusable is atomic feedback?

[RQ 1.4] How did teachers use and perceive the SA system?

[RQ 1.1] is of practical relevance: given that teachers complain about having too many correcting tasks (Eurydice, 2021; Gibson et al., 2015), we want to investigate the effect of the SA system on possible time savings and the amount of feedback compared to PP. The link between time investment and the amount of feedback stems from a vast body of literature stating that teachers often cope with time constraints in PP settings by limiting the amount of feedback by going into less detail (Junqueira & Payant, 2015; Price et al., 2010). Thus, our hypothesis is that possible efficiency gains of the SA system may be both in terms of time investments or amount of feedback. Nevertheless, more feedback does not necessarily mean better feedback (Glover & Brown, 2006; Higgins et al., 2001). However, this chapter does not investigate the content quality of the feedback in both conditions, nor issues related to the effectiveness of formative or summative feedback (Hattie & Timperley, 2007).

The whole idea of this study was based on a very broad idea of atomic feedback, presented below ([section 1.2](#)). [RQ 1.3] answers the usefulness of atomic feedback in light of reusing feedback. Before [RQ 1.3] can be answered, however, the broad idea of atomic feedback must be narrowed down to content-specific guidelines, making it possible to distinguish feedback items into atomic/non-atomic, which is done in [RQ 1.2]. The need for content-specific guidelines to make that distinction came during the data-analysis process: the broad idea of atomic feedback ([section 1.2](#)) was not sufficient to get high inter-rater reliability estimates between two blind, independent coders making the distinction. We will discuss this further in the data analysis procedures ([section 1.3.3.1](#)).

[RQ 1.4] wants to see how certain teachers' characteristics influence the use of the SA technology: do school level, gender, age and computer skills matter when interacting with the SA system? In addition, the teachers' perceptions of the SA-system are investigated: if we want teachers to embrace SA as a useful technology for practice, not only time savings or amount of feedback are important factors to look at. The Technology Acceptance Model TAM (Davis et al., 1989; see Figure 1.1) describes a pathway in which personal beliefs shape the attitude towards using a technology, which in its turn shapes intentions to use it. In other words, also teachers' views do matter. Two critical beliefs deserve special attention: (1) perceived usefulness and (2) perceived ease of use. Perceived usefulness is the user's subjective belief that using a technology will increase his/her performance and productivity. Perceived ease of use refers to the user's belief that the use of a technology will not take much effort. Several studies have shown that perceived usefulness and perceived ease of use are significant predictors of attitude towards technology use and intention to use it (e.g. Anni et al., 2018)

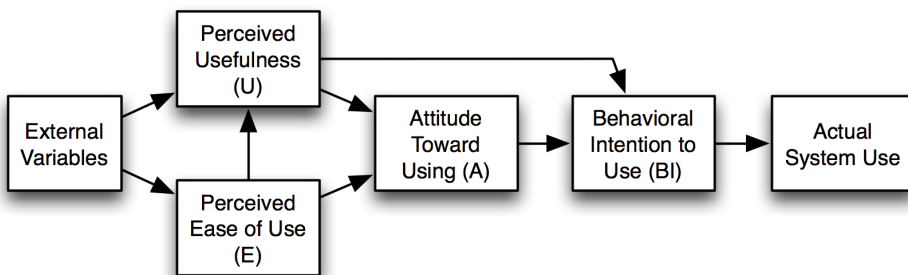


Figure 1.1 – The TAM-model (Davis et al., 1989)

1.2 ATOMIC FEEDBACK

1.2.1 Definition

Classic written feedback has traditionally consisted of long pieces of written text (Winstone et al., 2017). In the SA system, feedback is saved for easy reuse with other students making the same or similar mistakes. With its long sentences describing all of the errors in a student's work, classic written feedback is intrinsically not very reusable, as it is explicitly targeted toward specific students (Glover & Brown, 2006). To overcome this difficulty and maximize the reusability of feedback, one of the key ideas underlying the proposed SA system is that it promotes to teachers the writing of *atomic feedback*.

Atomic feedback (see Figure 1.2) counters the need to write long pieces of feedback that describe many different mistakes at once. In atomic feedback, a teacher must:

1. identify independent errors,
2. write small feedback items for each error separately, and

3. if an error reflects a structural mistake or misconception (Gusukuma et al., 2018), the teacher must create two feedback items:
 - (a) one item containing feedback on the misconception in general and
 - (b) one or more sub-items addressing specific mistakes.

Atomic items ultimately form a point-by-point list covering only items that are relevant to a student's solution. The list can be hierarchical, in order to *cluster* items that belong together. Clustering ensures that feedback can be written as atomically as possible and prevents teachers from writing overly specific items because it provides an orderly way in which to present related feedback to students (e.g., through thematic clustering or a visual presentation of both general and specific feedback on the same error).

A comparison of *classic* and *atomic* feedback is presented in Figure 1.2. As demonstrated by this comprehensive example, any classic feedback text can be rephrased as an atomic text. All the items are examples of atomic feedback items; the list hierarchy is first dedicated to the step in the student's solution and second to the kind of errors. The structural mistake with making the denominators the same, leads to two feedback items: one announcing the structural mistake (concerning making the denominators the same), one on the specific error (' $2\pi r^2$ shouldn't be directly in the numerator').

Student's solution
Manipulate the formula: $A = 2\pi rh + 2\pi r^2$ to h

$$\frac{A}{2\pi \cdot r} = h + 2\pi r^2$$

$$\frac{A - 2\pi r^2}{2\pi \cdot r} = h$$

<p>Classic feedback</p> <p>Mind the fact that the dominant operation on the right-hand side of the equation is an addition! The division of the left-hand side by $2\pi r$ is, therefore, not helpful. Moreover, $2\pi r$ is a common factor of the right-hand side, but the sum wasn't completely divided by it (second addend not divided). Although your final answer is correct, the way it is written makes it look like a coincidence. Going from the first to the second step, you would normally subtract $2\pi r^2$ from both sides, meaning that it shouldn't be placed directly in the numerator, as you should make the denominators the same.</p>	<p>Atomic feedback</p> <ul style="list-style-type: none"> • First step <ul style="list-style-type: none"> – Dominant operation on the right side is an addition! * Division of left-hand side is not helpful * $2\pi r$ is a common factor of the right side, but: <ul style="list-style-type: none"> · sum wasn't completely divided by it · the second addend was not divided • Second-step <ul style="list-style-type: none"> – Your final answer is correct, but: <ul style="list-style-type: none"> * It looks like a coincidence. * You should subtract $2\pi r^2$ from both sides. * Mistake with making the denominators the same! <ul style="list-style-type: none"> · $2\pi r^2$ shouldn't be directly in the numerator.
--	---

Figure 1.2 – A comparison of classic and atomic feedback

1.2.2 Hypothetical division into sub-items

To determine whether a feedback item is atomic or not, a teacher can try to divide the item into sub-items. If such hypothetical division makes sense with regard to the task being corrected, the item is non-atomic. For example, the feedback text *'Neither the choice of the unknown nor the starting equation is correct.'* can easily be divided into the sub-items *'Choice of the unknown incorrect'* and *'Initial equation incorrect.'* The advantage of dividing the item into two separate sub-items is that the sub-items can be reused independently for a potentially much larger group of students: those who only got the unknown wrong, those who got only the initial equation wrong, and those who got both wrong.

In some cases, however, a hypothetical division into sub-items might not feel meaningful, or it might even change the tone of the feedback. In that case, an item is still considered atomic. For example, the item *'What is your reasoning? I don't see any solution steps.'* could hypothetically be divided into *'What is your reasoning?'* and *'No solution steps shown.'* However, such a division would not make sense, if these sub-items are always used together. The original item may thus be viewed as atomic. Another example is the following: *'Final answer is correct due to a combination of errors.'* This item could be split into *'Final answer is correct.'* and *'Combination of errors.'* However, by emphasising that the final answer is correct, the hypothetical division into sub-items changes the tone of the feedback, thus suggesting that the original item is indeed atomic.

1.2.3 Atomic or not?

In addition to the hypothetical division in sub-items, other violations of the definition of atomic feedback can also occur to render an item non-atomic. For instance, the independence condition can be violated if a feedback item refers to other items, thus making it impossible to use them separately. For example: *'Idem to the comment above'* cannot be used without the 'comment above.' In addition, counting the number of times an error occurred (e.g., *'Adding fractions with unlike denominators results in two errors'*) violates the independence condition of atomic feedback too. It would be much better to create a single main item —*'Adding fractions with unlike denominators'* (a common misconception)— with each specific error as a sub-item.

In summary, whether a feedback item is or is not atomic can be determined by dividing it into hypothetical sub-items or by assessing it according to the definition of atomic feedback. The extent to which a hypothetical division into sub-items is meaningful is often contextual and subjective.

1.3 MATERIALS & METHODS

In this section, we first present the designed materials for the experiment: the SA-tool in which feedback could be reused, the task on linear equations teachers had to assess, and the survey used to query the teacher after the experiment. Next, in the 'Methods'-part we thoroughly explained the study design and how these materials were used throughout the study.

1.3.1 Materials

1.3.1.1 Developing the SA tool

After formulating the concept of atomic feedback, we developed an initial version of the SA tool with reusable feedback. It was developed as an advanced grading method plug-in for Moodle, an open-source e-learning platform (Gamage et al., 2022). The choice to develop the tool in Moodle instead of developing an independent tool was based on several design-related reasons. Firstly, developing the SA tool as a plug-in for the world-famous Moodle platform could offer high international exposure through open-source publication. Secondly, the Moodle framework could accelerate the development of the application, as it provides ready access to many necessary components (e.g., gradebook, assignment uploading, login).

The SA tool makes it possible to formulate feedback items in form of a hierarchical list, as shown in [Figure 1.3c](#). Many keyboard shortcuts were implemented to indent, delete, and reshuffle items quickly. Items are saved for reuse with subsequent students, and the suggestion system tries to match what a teacher is typing with previous feedback items.

1.3.1.2 Task on linear equations

For the crossover study, we also developed a task on linear equations, in close cooperation with a ninth-grade mathematics teacher. This topic was especially appropriate for the experiment, as it was likely to be familiar to all math teachers in Flanders (the Dutch-speaking part of Belgium). The task consisted of three items: (1) solving an equation (easy/procedural), (2) manipulating a formula (complex/procedural; see [Figure 1.2](#)), and (3) a modeling question consisting of a word problem (complex/problem-solving). The three items were combined to form a representative, standard task on linear equations. The task and its model solution are presented in [Appendix A](#).

1.3.1.3 Teacher survey after the experiment

Teachers were asked to fill in a survey after the experiment. The questionnaire contained three parts surveying:

1. Some personal information (age, teaching experience, school types, and grade they were teaching),
2. Their self-reported computer skills,
3. Their view of the SA system.

To measure their self-reported computer skills (part 2), we used the 30 validated items of the Computer User Self-Efficacy-scale, abbreviated as the CUSE-scale, developed by Cassidy and Eachus (2002). Their scoring procedure leads to an overall CUSE score ranging from 30 to 180 for every teacher; the higher the CUSE score, the more positive their computer self-efficacy beliefs. Part 3 contained 12 items, measured on a 7-point Likert scale based on the Technology Acceptance Model (TAM) (Davis, 1989; Davis et al., 1989; Venkatesh et al., 2003). The items polled teachers' perceived usefulness, perceived ease of use, attitude and behavioral intention to use the SA-system. This part of the survey can be found in [Appendix C](#).

1.3.2 Methods

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

1.3.2.1 Participants

Baseline characteristic	Pilot study		Actual study		Full sample	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Total	9	100	36	100	45	100
Gender						
Female	6	66.7	22	61.1	28	62.2
Male	3	33.3	14	38.9	17	37.8
Teaching experience						
3–5 years	4	44.4	7	19.4	11	24.4
5–10 years	2	22.2	4	11.1	6	13.3
10–20 years	2	22.2	14	38.9	16	35.6
20–30 years	1	11.1	6	16.7	7	15.6
> 30 years	0	0	5	13.9	5	11.1
Upper/Lower secondary education^b						
Lower secondary education	5	55.6	16	44.4	21	46.7
Upper secondary education	4	44.4	20	55.6	24	53.3
School type^a						
General secondary education	7	77.8	31	86.1	38	84.4
Artistic secondary education	0	0	1	2.8	1	2.2
Technical secondary education	5	55.5	15	41.6	20	44.4
Vocational secondary education	0	0	1	2.8	1	2.2
Grades taught^a						
7 th grade (12/13-year-olds)	3	33.3	7	19.4	10	22.2
8 th grade (13/14-year-olds)	3	33.3	6	16.7	9	20.0
9 th grade (14/15-year-olds)	4	44.4	14	38.9	18	40.0
10 th grade (15/16-year-olds)	6	66.6	16	44.4	22	48.9
11 th grade (16/17-year-olds)	4	44.4	17	47.2	21	46.7
12 th grade (17/18-year-olds)	4	44.4	18	50.0	22	48.9

Note. The average age of the participating teachers was 40.2 years ($SD = 10.3$) for the full sample, 33.4 ± 6.8 years for the pilot study and 41.8 ± 10.4 years for the actual study.

^aSome teachers are counted multiple times, as they taught several different classes.

^bTeachers are upper secondary education teachers if they teach at least one 11th or 12th grade class.

Table 1.1 – Characteristics of the participating mathematics teachers

60 ninth-grade students in one secondary school in Flanders (Belgium) solved the task on linear equations. All of the students were 14 or 15 years of age, and they all were taking math with the teacher who had helped to develop the task ([Appendix A](#)).

The study was announced in three mathematics education magazines in Flanders, as well as on Facebook groups for Flemish math teachers. As a result, 45 secondary mathematics teachers with at least three years of working experience from all parts of Flanders volunteered to participate in the study, which was conducted in the summer of 2020. The teachers were reimbursed for their travel expenses and offered lunch on the day of the study. Given the manner of participant recruitment, more committed math teachers may be overrepresented in the sample. As shown in [Table 1.1](#), however, we did manage to assemble a diverse group of teachers in terms of gender, age, school type, grades taught, and teaching experience.

The first time we organized the experiment with 9 math teachers, we noticed several methodological imperfections in our study design. We decided to use this first try as a pilot study ($n = 9$) to refine the actual study ($n = 36$).

1.3.3 Study design

In this part we thoroughly explain the study design. Some design choices were made due to experiences in the pilot study. The design of the pilot study and the reconsiderations it led to, can be found in [Appendix B](#).

Before the experiment with teachers, 60 students had completed the math task on linear equations as a test during an actual math class (i.e., an authentic context) in February 2020. The students had been studying linear equations in class, their teacher had developed the task, and the grades were later incorporated into the students' grade reports.

The experiment with teachers was a crossover study with two conditions (Bose & Dey, 2009): SA and PP. In the SA condition, teachers were asked to use the SA tool to write atomic feedback items. When a teacher started typing, the system searched for matching feedback items to reuse (see [Figure 1.3c](#)). In the PP condition, teachers received only a text field in which to type their feedback (see [Figure 1.3d](#)). Perhaps surprisingly, the PP condition did not involve any 'real' paper-based correction work, as had been the case for the pilot study (see [Appendix B](#)). Results from the pilot study indicated that such a design actually investigated whether teachers wrote faster on a computer or on paper, and not the added value of reusable feedback.

Each teacher assessed solutions to all 60 tasks, with a quasi-random selection of 30 tasks being assessed under the SA condition and the other 30 tasks under the PP condition. The computer algorithm used to draw this quasi-random selection ensured that, throughout the entire study, each task would overall be assessed the same number of times under both conditions and that each condition would contain 10 tasks with good answers, 10 with moderate answers, and 10 with poor answers. This task distribution was necessary in order to allow comparison of the two conditions for each teacher (i.e., it avoided accidentally over-representing tasks with good answers in one condition and over-representing those with poor answers in the other) because imbalance between conditions could potentially affect the time spent and amount of feedback provided in each condition, thereby biasing the results of [\[RQ 1.1\]](#). To classify the 60 tasks as good,

Start correcting Question 1

Question 1 must be corrected

Start correcting Question 2

Question 2 must be corrected

Start correcting Question 3

Question 3 must be corrected

2. Manipulate the formula to h

$$\frac{A}{2 \cdot \pi \cdot r} = n + 2 \pi r^2$$

$$\frac{A - 2 \pi r^2}{2 \cdot \pi \cdot r} = n$$

Atomic feedback Perfect (Max) Not answered (0)

2 /10

Stop correcting Question 2

a. b.

2. Manipulate the formula to h

$$\frac{A}{2 \cdot \pi \cdot r} = n + 2 \pi r^2$$

$$\frac{A - 2 \pi r^2}{2 \cdot \pi \cdot r} = n$$

- First step
 - Dominant operation on the right-hand side is an addition!
 - Division of the left-hand side is not helpful
 - $2\pi r$ is a common factor of the right-hand-side, but:
 - sum wasn't completely divided by it
 - second
 - the second addend was not divided

Calculated grade: 2/10

2 /10

Stop correcting Question 2

c. d.

2. Manipulate the formula to h

$$\frac{A}{2 \cdot \pi \cdot r} = n + 2 \pi r^2$$

$$\frac{A - 2 \pi r^2}{2 \cdot \pi \cdot r} = n$$

Your feedback:

Mind the fact that the dominant operation on the right-hand side of the equation is an addition! The division of the left-hand side by $2\pi r$ is, therefore, not helpful. Moreover, $2\pi r$ is a common factor of the right-hand side, but the sum wasn't completely divided by it (second addend not divided). Although your final answer is correct, the way it is written makes it look like a coincidence. Going from the first to the second step, you would normally subtract $2\pi r^2$ from both sides, meaning that it

2 /10

Stop correcting Question 2

Figure 1.3 – SA tool integrated into Moodle (Screens a–c) and the screen for the PP condition (Screen d)

moderate, or poor, we used the average grades given by the teachers in the pilot study (see Appendix B): the 20 tasks with the highest grades were rated as good, the next 20 as moderated, and the 20 lowest grades as poor.

To mitigate order effects inherent in crossover experiments (Bose & Dey, 2009), half of the teachers started under the SA condition, and the other half started under the PP condition. It was further necessary to control for bias emerging from increasing task familiarity (Lim et al., 1996) under both conditions. In both conditions, the 30 tasks were therefore presented to the teachers in random order. This procedure ensured that developments over time – an aspect essential to investigate [RQ 1.1] – did not depend on the order in which the tasks were assessed. For example, if all of the ‘good’ tasks were presented at the end, the teacher’s time investment for each task would suddenly decrease, but these time savings would not be due to the condition but to the quality of the responses to the tasks.

The experiment was executed in the summer of 2020. Unfortunately, due to COVID-19 measures, only nine participants could be received in the computer lab each day. We

therefore repeated the experiment seven times throughout the summer holidays.

Each day started with a training for all participating teachers. Aside from the practical arrangements, the training focused largely on working with the SA tool and the PP-text field in Moodle, as well as on how to formulate atomic feedback. To avoid influencing the teachers, they were not informed about the research questions. The linear equation task was also never mentioned during the training, and a geometry task obtained from other students was used instead as a demonstration. Teachers received as much time as they deemed necessary to practice using the software, thus avoiding bias due to learning effects. At the end of the training, teachers were asked to treat the students' tasks in the experiment in the same way that they would treat their own students' tasks (no questions about how to assess certain student solutions were allowed) and to aim to provide the same quality of feedback in both conditions.

After the training, the solution key for the linear equation task was distributed (see [Appendix A](#)). The teachers were asked to review it in detail and note which feedback they would formulate. Once they had done so, the experiment started in either the SA or PP condition. They received maximum 2.5 hours to finish assessing the 30 tasks in the starting condition. After the first condition was finished, we had a lunch break, after which the teachers again received 2.5 hours for the 30 tasks remaining in the other condition.

To measure the time spent on each item, teachers pressed a 'Start' button each time they started assessing a question and a 'Stop' button when they were finished (see [Figure 1.3a/1.3b](#)). This procedure was the same in both conditions. When no question was being assessed, no time was registered. Once they started, the teachers were asked to work through the student's solution in one go, pausing only when no question was being assessed. This time-registration technique was deliberately chosen in favor of more automated time registration approaches (screen recording, counting the time a teacher spent on one student,...) as teachers can be distracted during correction work (for example, by receiving a text message), which would introduce a bias in the data. Using this time recording mechanism, teachers knew that pressing 'Start' required assessing a student's answer without interruptions. However, it had some drawback as well: once an item was assessed, it was not possible to return to it (this would have made it very hard to register the total time correctly), and it required a teacher to correct the three questions of the linear equation task for each student consecutively. For some teachers, this felt odd as they did not usually assess student-by-student (i.e., assessing all responses of a student at once), but rather preferred a question-by-question approach (i.e., assessing all responses to a question at once) (McMillan, 2013).

For the SA condition, teachers always had three options: formulating atomic feedback, indicating that a solution was perfect, or indicating that a solution was missing (see [Figure 1.3b](#)). The buttons with the pre-defined atomic feedback 'Perfect' or 'Not answered' were mainly introduced to give a demo of the system, as Moodle complains when not all question are corrected. When formulating feedback, they could use keyboard shortcuts (e.g., to indent, insert, or remove items) to create a hierarchical list of feedback items (see [Figure 1.3c](#)). When a teacher typed something, the system searched the feedback items that had already been entered to detect possible matches (see [Figure 1.3c](#)). Thus, the suggestion system was non-intelligent. The system searched only within the feedback

items that the teacher had already entered for that particular question. The main reason for this design choice was that we currently do not know if it is practically feasible to share feedback among teachers: writing styles, explanations, priorities,... can differ among them. The same issue is at play to a lesser extent when it comes to letting teachers share their feedback items across questions: at this moment, we do not know if feedback items could remain the same across questions since their phrasing, importance, and prominence can differ (see [Section 1.6, Further research](#)). Moreover, the non-intelligent suggestion system would probably have made the sharing of feedback items across teachers and questions confusing.

In the PP condition, the teachers received only a text box in which to type feedback (see [Figure 1.3d](#)), with no possibility of reusing feedback. The buttons ‘Perfect’ or ‘Not answered’ were not available in this condition, as teacher can neither use such buttons when giving feedback using paper-and-pencil.

In the crossover study, the teachers were also asked to grade every question in both conditions. The process of setting up a grading system in conjunction with atomic feedback exceeds the scope of this chapter.

1.3.3.1 Data analysis procedures

We only used the data from the actual study ($n = 36$). The data from the pilot study were omitted, due to excessive differences in study design (see [Appendix B](#)). All statistical analyses were performed using R version 4.3.

Paired t -test were used to compare time differences and differences in amount of feedback ([\[RQ 1.1\]](#)). The amount of feedback was expressed as the total number of characters used by a teacher in one condition. The pairs consisted of the outcomes of each teacher in the two conditions. Data imputation was applied, as some time registrations were unreliable (e.g., one teacher accidentally started a question twice). At the end of each task in Moodle, teachers could explain anything that had gone wrong (e.g., ‘*reopened because I forgot to give feedback and Moodle complained.*’), thus allowing for reliable time registrations in most cases. For cases with multiple time registrations that were not explained and in which the erroneous registration was not clear, the mean of the multiple registrations was substituted. Data on the amount of feedback were only missing if a teacher had immediately scrolled to the next student without pressing the ‘Stop’ button for the preceding question. The teachers had been warned that, if that happened, the feedback for the previous question would be lost. In the few cases that did occur (22 in a total of 3,195 corrections), the mean of the number of characters used by the teacher for that question in that condition was substituted. No data outliers were removed, as all extreme observations corresponded to deliberate feedback styles of certain teachers.

For [\[RQ 1.2\]](#), we coded every feedback item written by the teachers in the SA condition as either atomic or non-atomic. To ensure the reliability of the process, we recruited a student assistant from the mathematics department of the University of Antwerp to perform the same coding task blindly and independently. We started with the idea that the general idea of atomic feedback, presented in [section 1.2](#), would be sufficient to code all items with high inter-rater reliability. After coding the first 1000 items, we concluded that this approach was too naive and that exact, content-specific guidelines are needed

to distinguish atomic from non-atomic ones. As a result, [RQ 1.2] also became a research question in itself, which was not initially intended. Hence, we planned multiple coding iterations with discussions about the differences between the raters at the end of each iteration to arrive at a comprehensive set of content-specific guidelines for atomic feedback. We aimed for an almost unambiguous codebook (Appendix D) and a Cohen's κ at least 0.8, implying high inter-rater reliability (Landis & Koch, 1977). After the final iteration, we investigated the items identified as non-atomic by both raters. We reviewed for each non-atomic item the applicable violations of atomic feedback from the developed codebook (see Appendix D). In hindsight, it would have been better if we had developed the codebook based on the pilot study data; and then used it on the actual data. However, due to the surprising and unexpected need for [RQ 1.2], we have overlooked this.

A Chi-square test of independence was used to analyse the relationship between the reusability and 'atomicness' of feedback items [RQ 1.3] in the SA condition. An item was considered reused if it was used more than once in the SA condition, and considered as (non)-atomic if it was categorised as such by both raters for [RQ 1.2].

Since [RQ 1.4] on teachers' usage is more about exploring which teacher characteristics influenced the use of the SA system, without prior hypotheses, we only used descriptive statistics. We correlated teacher characteristics with what they did during the experiment in both conditions. For teachers' views, we reported the outcomes on the different scales of the TAM model and their correlations. Nowadays, it is commonplace to report TAM as a structural equation model (Scherer et al., 2019), but this is beyond the scope of this study as it would have required an unfeasible number of participants given the experimental setup.

1.4 RESULTS

1.4.1 Differences in time and amount of feedback under the two conditions [RQ 1.1]

To compare time differences, we first calculated the total amount of time that a teacher spent assessing a task. Recall that teachers were asked to press a Start/Stop button when they wanted to start/stop giving feedback on a question, so that the time could be tracked precisely. Because each task consisted of three questions, we summed the time spent on each question to obtain the total time for a task. The average time needed to assess a task under the PP condition was 191 seconds (3 min 11 sec), with an average of 198 seconds (3 min 18 sec) under the SA condition.

To determine how the amount of time developed as more tasks were assessed, we averaged the time of all tasks assessed at the same instant (i.e., all tasks that were assessed first, second, third, etc.) separately for each condition. The tasks were presented to teachers in unique distributions across the two conditions, with a random running order to assess them. Each task was thus presented to each teacher at a different position, and we can therefore assume that a relatively equal amount of good/moderate/poor tasks were assessed at each instant, excluding potential biases. According to the results (Figure 1.4), time decreased greatly as more tasks were assessed in both conditions. The assessment of the first task took 367 seconds (6 min 7 sec) in the SA condition and



Figure 1.4 – Evolution of teachers' assessment time in both conditions as more and more tasks get assessed

276 seconds (4 min 36 sec) in the PP condition. For the 30th and final task, the time decreased to 128 seconds (2 min 8 sec) for the SA condition and 136 seconds (2 min 16 sec) for the PP condition. The bumpiness of both graphs stems from several instants containing slightly more good/poor tasks, which can never be avoided entirely.

A paired t -test was conducted to compare the total time (in seconds) that teachers spent assessing the 30 tasks in the SA condition and the 30 tasks in the PP condition. All preliminary assumptions for a paired t -test were met: (1) all teachers are independent subjects, (2) the total time used in each condition for each teacher was compared, clearly indicating paired samples, and (3) the Shapiro-Wilk-test indicated a non-significant deviation from normality of the differences ($p=.281$), so the differences are approximately normally distributed. The difference between the SA condition ($M = 6017 \text{ sec} = 1\text{h } 40 \text{ min } 17 \text{ sec}$, $SD = 1687 \text{ sec} = 28 \text{ min } 7 \text{ sec}$) and the PP condition ($M = 6016 \text{ sec} = 1\text{h } 40 \text{ min } 16 \text{ sec}$, $SD = 2180 \text{ sec} = 36 \text{ min } 20 \text{ sec}$); $t(35)=0.002$, $p=.998$ with 95% CI [-634, 636] was not significant. This could also be inferred intuitively from the two graphs in [Figure 1.4](#), as both exhibit almost equally steep downward trends.

The next step involved assessing the amount of feedback in both conditions. We express the amount of feedback as the total number of characters that each teacher used in each condition. The boxplots of the number of characters in each condition are presented in

Figure 1.5. A paired t -test was conducted to compare the amount of feedback characters in each condition. All preliminary assumptions for this paired t -test were met: (1) all teachers are independent subjects, (2) the total number of feedback characters in each condition for each teacher was compared, clearly indicating paired measures, and (3) the Shapiro-Wilk-test indicated a non-significant deviation from normality of the differences ($p=.191$), so the differences are approximately normally distributed. The paired t -test revealed a significant increase in the average number of characters in the SA condition ($M = 9656$ chars, $SD = 3553$ chars) relative to the PP condition ($M = 8409$ chars, $SD = 3672$ chars); $t(35) = 2.43$, $p = .02$ with 95% CI [207, 2288]. The effect size for this analysis ($d = 0.41$) approached Cohen’s convention (1988) for a medium effect ($d = 0.5$).

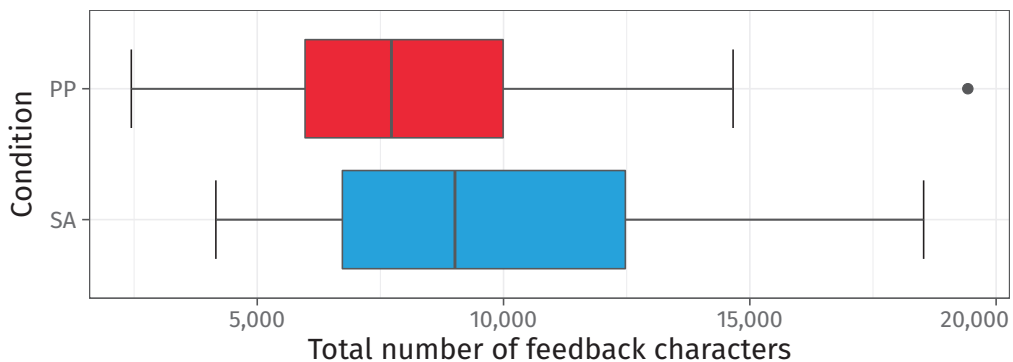


Figure 1.5 – Boxplots of the total number of feedback characters used by teachers in each condition

1.4.2 Content-specific guidelines for atomic feedback & measurement of the atomicness of the given feedback [RQ 1.2]

To determine the relative ‘atomicness’ of the feedback formulated in the SA condition of the crossover study, we analysed all 2,591 feedback items together with a student assistant from the mathematics department. Four coding iterations were needed to achieve profound content-specific guidelines of atomic feedback with a Cohen’s κ value of 0.84, which is generally accepted as almost perfect agreement (Landis & Koch, 1977). In the first iteration, we used only the theoretical definition of atomic feedback (see section 1.2) and only a vague notion of a hypothetical division into sub-items (‘If you can divide the item in your head, it is not atomic’). After categorising the first 1,000 feedback items, inter-rater agreement was moderate, with a Cohen’s κ value of 0.53. After the first iteration, systematic differences in coding were investigated and discussed thoroughly, leading to the need for a codebook that enhanced the definition and stated more precisely when hypothetical division into sub-items would or would not be meaningful, thereby improving the content-related and context-specific guidelines. To determine whether this enhanced codebook would succeed in achieving good agreement, we first tried it with 100 random items in the second iteration. We then made several additional minor adjustments to the codebook for the third iteration and determined the coding for the entire dataset. After the third iteration, we reached a substantial Cohen’s κ value of 0.66. Differences in the coding nevertheless clearly revealed the persistence of several systematic differences that had not been incorporated into the codebook. We

therefore adjusted the codebook again, recoded the items, and achieved a Cohen's κ value of 0.84. The results of all iterations are presented in [Table 1.2](#).

Iteration	Cohen's κ	% agreement	Level of agreement	Coded items
1	0.53	81.2	Moderate	First 1000 items
2	0.70	88.0	Substantial	100 random items
3	0.66	86.8	Substantial	Full dataset (2,591 items)
4	0.84	93.5	Almost perfect	Full dataset (2,591 items)

Table 1.2 – Summary table of inter-rater reliability in the various iterations

The final codebook (with references to the iterations in which guidelines were added or adjusted) is included in [Appendix D](#).

After the fourth iteration, no agreement was found for 169 feedback items (6.5% of 2591 items) concerning whether they were or were not atomic. Of the items that were coded in agreement, 1,784 (73.7%) were identified as atomic, with 638 (26.3%) coded as not atomic (638 items). The numbers of violations of atomicness occurring in the non-atomic items (based on the codebook in [Appendix D](#)) are listed in [Table 1.3](#). The table contains double counting, as some feedback items violated multiple guidelines.

Violation	No. of occurrences	% occurrences
The item contains both a comment/mistake and the location where the comment/mistake occurred.	272	37.7
The item discusses multiple errors/issues/remarks.	230	31.9
The item contains references to both a structural error/misconception and the specific mistake.	95	13.2
The item contrasts an error/remark to the entire solution process.	53	7.3
The item makes links between solution steps.	38	5.3
The item contains a reference to the number of times a mistake occurred.	20	2.8
The item makes an avoidable reference to another item.	14	1.9

Table 1.3 – Overview of the frequency of violations of atomicness in the 638 non-atomic feedback items

1.4.3 Reusability of atomic feedback [RQ 1.3]

A Chi-square test of independence was performed to examine the relationship between the atomicness of the feedback items and their reusability. An item was considered 'reused' if it was used more than once. An item was considered '(non-)atomic' if it was identified as such by both raters for [\[RQ 1.2\]](#). The crosstab is displayed in [Table 1.4](#). The relationship between these variables was significant: atomic items were more likely than non-atomic items to be reused, $\chi^2(1, n = 2424) = 85.34, p < .001$ with an odds ratio of 2.682 with 95% CI [2.165, 3.322].

Atomicness	Reusability		Total
	Reused	Not reused	
Atomic item	731 (40.9%)	1,055 (59.1%)	1,786 (73.7%)
Non-atomic item	131 (20.5%)	507 (79.5%)	638 (26.3%)
Total	862 (36.6%)	1,562 (64.4%)	2,424 (100%)

Table 1.4 – Crosstab for comparing the atomicness of feedback items that were or were not reused

1.4.4 Teachers’ usage & view of the SA system [RQ 1.4]

For the teachers’ usage of the SA-system, a correlation table is presented in Table 1.5 correlating different teachers’ characteristics (CUSE score, gender, age, teaching upper/lower secondary level) with the total time and total number of feedback characters

Variable	M ± SD	1	2	3	4	5	6	7	8	9	10
Teacher characteristics											
1. Computer self-efficacy (CUSE)	143.25 ± 15.41	—									
2. Gender ^a	17 male, 16 female	.43	—								
3. Age	41.83 ± 10.43	-.30	-.20	—							
4. Upper/lower secondary ^b	20 upper, 16 lower	-.08	-.20 ^c	.15	—						
SA condition											
5. Total time used	6017.18 ± 1686.73	-.20	-.10	.17	-.09	—					
6. Total feedback chars used	9656.50 ± 3552.81	-.08	-.08	-.29	-.37	.44	—				
7. Total number of feedback items	67.39 ± 29.23	-.16	-.19	-.15	-.08	.51	.74	—			
8. Atomicness of the items	74.68% ± 14.93%	.29	.23	-.13	.09	-.04	-.09	-.18	—		
9. Reusability of the items	39.09% ± 15.63%	.20	.21	-.13	.13	-.24	-.22	-.53	.68	—	
PP condition											
10. Total time used	6016.42 ± 2180.38	-.33	-.13	.24	-.11	.55	.12	.12	.10	.00	—
11. Total feedback chars used	8408.67 ± 3672.40	-.22	-.10	-.25	-.22	.12	.64	.54	-.04	-.21	.09

^a Point-Biserial correlation (1 = Male / 0 = Female)
^b Point-Biserial correlation (1 = Upper / 0 = Lower)
^c Phi correlation
 others: (regular) Pearson correlation.

Table 1.5 – Correlation table for teachers’ usage

used in both conditions. For the SA condition, the total number of feedback items is also reported, as well as the atomicness and reusability of the items: for each teacher, we calculated what percentage of their feedback items were coded as atomic (by both raters), based on the categorisation for [RQ 1.2], and for the reusability score we computed the percentage of their feedback items they had reused (at least once).

The teachers views are reported in Table 1.6. The table contains the mean, standard deviation, Cronbach's α and correlations of the scales stemming from the TAM model. The scales were calculated by averaging the corresponding items responses on a 7-point Likert scale. The corresponding items and their mean and standard deviation can be found in Appendix C. All scales reached a Cronbach's α higher than 0.7, which is generally accepted as a rule of thumb for scale reliability (Taber, 2018).

Scales	M \pm SD	Cronbach's α	1	2	3
1. Perceived Usefulness	5.09 \pm 1.05	0.84	—		
2. Perceived Ease of Use	4.94 \pm 0.99	0.75	.61	—	
3. Attitude Towards Using	5.33 \pm 0.81	0.80	.83	.50	—
4. Behavioral Intention to Use	5.15 \pm 1.28	0.97	.87	.71	.73

Table 1.6 – Results of the TAM model

1.5 DISCUSSION

In this study, we propose a new way of assessing handwritten tasks by having teachers collaborate with a computer when giving feedback. In this semi-automated (SA) approach, all feedback items are saved for reuse when subsequent students make the same or similar mistakes. Regarding [RQ 1.1], we sought to validate this approach experimentally by examining possible time savings and differences in the amount of feedback, compared to the classic PP approach. To make the use of the SA system beneficial, teachers need to know how to formulate reusable feedback. Therefore, we developed the concept of atomic feedback: a collection of form requirements for breaking down feedback into a hierarchical list of feedback items. Each item considers only one mistake at a time, and each item can be used independently. With regard to [RQ 1.2], we sought to investigate whether atomic feedback could be made distinguishable from non-atomic feedback using content-specific guidelines, whether teachers would be able to produce this type of feedback, and, if not, which violations of atomic feedback occurred. For [RQ 1.3], we investigated whether formulating atomic feedback can result in more reusable feedback. Finally, in [RQ 1.4] we looked into how teachers operated in both conditions, and whether this could be linked to personal characteristics. Moreover, teachers' views on the SA system were also examined.

For [RQ 1.1], our results indicate that, instead of saving time, the teachers in this sample tended to give significantly more feedback when using the SA tool. In other words: the participating teachers chose to provide more feedback instead of completing the correction job faster. Some teachers already had that feeling during the experiment: *"I think I'm giving much more feedback using your system"*, said one. Given our method of volunteer sampling, this result may be somewhat limited by selection bias. More

specifically, the sample consisted primarily of highly motivated mathematics teachers. The result was nevertheless unanticipated, as the teachers were asked to strive to provide comparable feedback in both conditions in terms of amount, quality, and content. This result also serves as a warning for anyone attempting to find ways of reducing the demands of the teaching profession. Although such solutions are of crucial importance (Eurydice, 2021), some teachers are likely to view them as opportunities to do even more work.

From a methodological perspective, our results serve as a reminder that one should be skeptical about claims that a given approach saves correction time while presenting time developments without any reference point. In our study, if we only had considered the SA graph in [Figure 1.4](#), the conclusion might have been that SA increases time savings as more tasks are assessed. Unless these results are compared to those of the PP condition, it is impossible to see that the downward trend is due to increasing task familiarity (Lim et al., 1996). Although we found no studies investigating task familiarity within the context of correction work done by teachers, it seems that teachers eventually memorise their solution keys and become routinised at giving feedback.

In this study, we found that atomic feedback could be distinguished ([\[RQ 1.2\]](#)) within the context of a linear equation task after four iterations, with a high Cohen's κ value (0.84). The fact that agreement could not be achieved for 169 feedback items indicates that higher inter-rater reliability would be quite difficult to obtain. Some feedback items were too poorly described to categorise them, while others reflected teacher intentions that were subject to multiple interpretations. For example, consider the item *'Your thinking is correct but you start with the wrong unknown variables.'* It is debatable whether dividing this item into *'Your thinking is correct!'* and *'Started with wrong unknown variables.'* would change the intentions of the teacher. In other words, it is unclear whether Guideline 7 on being atomic in the codebook ([Appendix D](#)) does or does not apply.

As is clear from the violations of atomicity (see [Table 1.3](#)), the most common violation was providing both the location as well as the error in the same feedback item (e.g., *'Calculation error in Step 2'*). This violation indicates that teachers feel a need to highlight the location of error in their feedback, as the SA system only allowed to formulate feedback below the solution. Indeed, *'I really want to point out where the student goes wrong!'* was a frequently heard comment during the experiment. A future version of the SA system could allow teachers to tap on the mistake and create feedback items at the location of an error, immediately removing the teachers' need to indicate where things went wrong in the written feedback. Moreover, the place of a mistake could give some valuable information about which feedback items to suggest for reuse to teachers (see [Section 1.6, Further research](#)). Finally, the struggle of indicating the location of a mistake existed in both the SA and PP condition. Still, teachers could have lost some time on figuring out how to structure such feedback in an atomic way which could have contributed to the non-significant time difference for [\[RQ 1.1\]](#).

It was rather surprising that addressing multiple issues at the same time was only the second major violation (and not the first), accounting for less than a third of all violations of atomicity (see [Table 1.3](#)). Together with the fact that 74% of the feedback items were identified as atomic (see [Table 1.4](#)), these results confirm that teachers can easily be trained to formulate atomic feedback. This percentage could probably

be even higher by showing teachers clear examples of possible violations and their atomic alternatives. The codebook in [Appendix D](#) could be a good starting point, as it conceptualises atomic feedback in more detail for tasks on linear equations.

The most interesting finding was that atomic feedback items were reused significantly more than non-atomic items ([\[RQ 1.3\]](#)). This suggests that atomic feedback embodies form requirements that lead to reusable feedback. The formulation of atomic feedback is thus easy for teachers to learn, and it increases the reusability of feedback items.

This finding is also supported by the correlation table in [Table 1.5](#) on how teachers operated in both conditions ([\[RQ 1.4\]](#)). All teacher characteristics correlated only weakly or not at all with what happened during the experiment, suggesting that these characteristics seem to play little to no role when providing feedback in the PP or SA condition. Contrary to expectations, even the computer self-efficacy score (CUSE score) showed only weak to no relationships with the measurements of the conditions, implying that teachers' computer skills make little to no difference in how they interact with the SA/PP system. A note of caution is due here since these computer skills are self-reported. Nevertheless, only the teachers' behavior during the experiment seemed to matter: the time spent in both conditions correlated strongly (0.55), meaning that fast teachers were generally fast in both conditions and slow ones remained slow. The same was true for the amount of feedback in both conditions (0.64): verbose teachers continued to be verbose in both conditions, concise ones stayed concise. So, indeed, the only strong and interesting correlation was yet again the one between the atomicity and reusability of teachers' feedback items (0.68) in the SA condition.

Upon reflection, it is worth considering whether the terms *atomic*, *reusable*, and *reused* can be used interchangeably in the ideal case. They can not. None of the three combinations (atomic/reused, atomic/reusable, and reused/reusable) are synonymous. First, there is an important difference between being an atomic item and being a reused item: an atomic item adheres to the definition and guidelines in this paper, but it depends on the teaching context whether the item will be reused too. Although mistakes in mathematics are often structural and appear many times, unique mistakes still occur in class groups, or a mistake might relate to a misconception, but the concrete error might be unique. Writing feedback in an atomic way guarantees that in both cases, most feedback items in the report to the student might be pre-existing or reused for following students, except for these unique feedback items. The idea of atomic feedback is, therefore, also valuable for these items as it tries to limit the need for writing new items to the unique ones when the database is already filled with many feedback items; without compromising the possibility of writing tailored feedback for a student. Second, atomic and reusable are not synonyms: atomic feedback consists of guidelines for teachers while writing feedback, demonstrating to make the feedback more reusable. However, you can only decide whether an item is reusable after writing it, which means the concept of 'reusable feedback' is not useful during the writing process. Conversely, it is even clearer: if a particular mistake — connected to a misconception — pops up many times in a class group, a teacher might be tempted to make one feedback item that addresses both the particular mistake as well as the misconception. While the feedback item will undoubtedly be reusable, it does not adhere to the definition of atomic feedback that advises making one item that points to the misconception and another sub-item that addresses the particular mistake. Lastly and highly related to the previous explanation, reusable and reused are not equal either: reused is a characteris-

tic of a feedback item that can be determined after the process of giving feedback by checking if an item is used more than once; reusable is a hypothetical characteristic of a feedback item, indicating an expectation that an item could appear in the feedback of more than one student.

Regarding teachers' views ([RQ 1.4]), the results in Table 1.6 suggest that teachers agreed that the SA-system is useful to them: they showed a strong attitude towards using the SA system and a strong behavioral intention to use it. Perceived usefulness was strongly correlated with the attitude towards using SA (0.83) and the behavioral intention to use (0.87); moreover, the behavioral intention to use correlated strongly with all other measures. Teachers rated the perceived ease of use of the SA system positively, but lower than all other measures. This lower score can be explained by some comments teachers made during the experiment: almost all teachers liked the SA idea and wanted to use it immediately (*"Can you please notify me when it is available, I need this!"*), however, many of them told the researcher that they tended to forget what they had written as feedback items, meaning that they sometimes could not find the right feedback item to reuse (remember the non-intelligent suggestion system). Some teachers said that they would have structured their feedback differently now that they had fully experienced the SA system. A few teachers said it would be practical to prepare some feedback items beforehand, as only filling the database of feedback items by assessing more and more students goes against their tradition of preparing correction schemes in advance. A part of the teachers complained that they did not like to assess student-by-student, but preferred assessing question-by-question (this is already possible but was not allowed during the experiment). Finally, some teachers mentioned that they would like to share their feedback items with their colleagues to further enhance the efficiency of the feedback process. All these remarks are essential indications for further research.

1.6 FURTHER RESEARCH

Even though the atomic items were reused significantly more than the non-atomic items, most of the atomic items (59.1%, see Table 1.4) were not reused. One possible explanation for this result is that some errors were not repeated in the solutions because each teacher corrected only 30 student tasks in the SA condition; higher proportions of reuse might be naturally expected in larger classes or when reusing feedback across subsequent school years. However, another reason could be due to the non-intelligent suggestion system of feedback items during the experiment. The system attempted only to match what teachers were typing to feedback items that had already been formulated. During the experiment, many teachers stated that they tended to forget how they had phrased some items. Therefore, they could not find the matching item with the suggestion system. That sometimes forced them to formulate already given feedback again instead of reusing feedback. This phenomenon was confirmed by the identification of nearly identical feedback items during the coding process for [RQ 1.2]. These results indicate that enhancing the intelligence of the suggestion system in terms of proposing potentially suitable feedback items is an essential issue for further research. As previously noted, one way to do this would be to allow teachers to indicate where an error has occurred in a handwritten solution. In addition to solving this need on the part of teachers, the location could also provide helpful information about which feedback items are appropriate. We hypothesise that feedback items will often

occur at more or less the same location (e.g., some mistakes are usually made at the beginning of a solution, others more at the end.). Searching for patterns of items that occur together could also provide a smarter suggestion system, possibly using machine learning techniques.

During this experiment, feedback was not shared among teachers or questions (see [1.3.3 Study design](#)). Investigating the feasibility of sharing feedback across teachers and questions is an important issue for further research, as it can further increase the efficiency of the feedback process. This was also suggested by some teachers during the experiment (see [1.5 Discussion](#)).

A primary restriction of this chapter is that we did not qualitatively analyse the content quality of the given feedback. This is a fundamental issue as feedback is directed at learners. Therefore, we plan a study to compare the feedback quality of atomic and classic feedback based on the data of this experiment (see [Chapter 2](#)). Also, the student's perspective is needed to get a complete picture of atomic feedback. In addition, it would be interesting to investigate the potential for frequency analysis based on which feedback items are used. That is, comparing work that shares a particular feedback item. Such analysis might give possible directions for further teaching. Moreover, every feedback item gives some information about a student's proficiency. This opens up possibilities to extensively monitor students' learning processes and apply adaptive differentiated instruction using Bayesian networks. A Bayesian network is a probabilistic graphical model of a student's proficiency (Almond et al., 2015).

Finally, although this study was set up within the context of mathematics, we are convinced that the ideas of semi-automated feedback with reusable, atomic feedback can be adapted to various school subjects. Given that the content-specific guidelines of atomic feedback will differ in other areas, future studies on the current topic are recommended.

CRedit authorship contribution statement

Filip Moons: Conceptualisation, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Funding acquisition. ● **Ellen Vandervieren:** Methodology, Supervision, Funding acquisition. ● **Jozef Colpaert:** Supervision, Writing – review & editing.

HIGHLIGHTS

- ✔ Feedback is compared when teachers could use a tool to reuse already given feedback ('statement banks') and when teachers could not use such a tool. The two feedback types were compared using text mining and qualitative analysis. Text mining identifies meaningful patterns and new insights in text using computer algorithms. Analysing feedback by combining text mining with qualitative techniques is a relatively new methodological approach for educational research.
- ✔ Word frequencies, sentiments and the amount of erroneous, descriptive and corrective feedback were similar in both feedback types. When teachers used the tool to reuse feedback, the feedback was more elaborate but less specific to the student's solution. Without the tool, feedback was shorter but more concrete and focused on the main issues. Overall, the tool to reuse feedback diverted teachers to less effective diagnostic activities.
- ✔ While a tool to reuse feedback can be helpful, it is important for teachers to avoid confusing handiness with quality. When using such a tool carelessly, our research suggests that some teachers may be inclined to simply describe and correct students' work rather than taking the time to analyse underlying misconceptions or misunderstandings. Therefore, it is crucial to continue paying attention to the quality of feedback, regardless of the feedback type.



☑ Chapter 2

COMPARING REUSABLE, ATOMIC FEEDBACK WITH CLASSIC FEEDBACK ON A LINEAR EQUATIONS TASK USING TEXT MINING AND QUALITATIVE TECHNIQUES

PUBLICATIONS

Moons, F., Holvoet, A., Klingbeil, K. & Vandervieren, E. (Under review). Comparing reusable, atomic feedback with classic feedback on a linear equations task using text mining and qualitative techniques. *British Journal of Educational Technology*.

Moons, F., Holvoet, A. & Vandervieren, E. (2022). Comparing atomic feedback with classic feedback on a linear equations task using text mining techniques. *Proceedings of the 13th ERME Topic Conference on Mathematics Education in the Digital Age (MEDA 3)*, 7-9 September 2022 in Nitra, Slovakia. <https://hal.science/hal-03925304>

ABSTRACT

In this crossover experiment, we examined how using a semi-automated tool that saves previously written feedback (SA condition) affected 45 math teachers' feedback on 60 completed linear equation tasks compared to writing 'classic' feedback as they would with pen-and-paper (PP condition). In the SA condition, teachers were encouraged to use atomic feedback, a set of formulation requirements that makes feedback items significantly more reusable. A previous study found that significantly more feedback was written in the SA condition but did not investigate the differences and similarities of the provided feedback. In order to fill this gap, we used text mining and qualitative techniques. Our results showed that while word frequencies, sentiments, and the amount of erroneous, descriptive, and corrective feedback were similar in both conditions, SA feedback was more elaborate but more general and focused on major and minor strengths and deficits. In contrast, PP feedback was shorter but more concrete and focused on the main issues. Our findings show that the feedback quality was low in both conditions, but using the semi-automated system diverted teachers to even less effective diagnostic activities. Furthermore, this study highlights how text-mining techniques can enhance qualitative research.

2.1 INTRODUCTION

Feedback has been recognised as a crucial component in learning processes (Busch et al., 2015a; Shute, 2008). While many studies in educational technology have considered the effect of the modes of delivery of feedback (e.g., Gleaves and Walker, 2013; Ryan et al., 2019) and the effect of immediate versus delayed feedback (e.g., Candel et al., 2020; Lefevre and Cox, 2017), in this chapter, we focus on the role of technology as it helps teachers to write feedback on a mathematics task. More specifically, we compare the written feedback reports composed by teachers under two conditions: the semi-automated condition (SA) in which they could use software to reuse previously written feedback items working like a statement bank (Denton & McIlroy, 2018; Moons et al., 2022) and the paper-and-pencil condition (PP) in which teachers could not reuse feedback, resembling regular feedback on a paper-and-pencil task (Chang et al., 2012), but instead of being handwritten, it is being typed.

ABBREVIATIONS • We will abbreviate the conditions as **SA** and **PP** in the rest of the chapter and also refer to them as feedback types.

Ideally, written feedback reports should strike a balance between the volume and focus on the main issues as more feedback does not necessarily mean better feedback (Glover & Brown, 2006). Indeed, Evans (2013) indicates that feedback should not be so specific and detailed that students do not have to think for themselves anymore. Chiles (2021)

calls this balance the ‘goldilocks principle’: feedback should be concise and accurate since “too much feedback can be overwhelming for students and lead them to disengage with it.” It seems to be best to link feedback directly to overarching learning intentions and break it into small, achievable steps. As such, feedback should be more than solely corrective: it should indicate the what, how and why of problems in the students’ work (Gibbs & Simpson, 2005), address misconceptions (Schnepper & McCoy, 2014; Yang & Lu, 2021) and identify actions the student can take to improve (Sadler, 2010).

Providing feedback may be tedious and time-consuming: 49% of the teachers in the European Union and 53% of all British teachers complain about having too much assessment work (Eurydice, 2021; Gibson et al., 2015). One of the well-known coping mechanisms to overcome this workload is shortening feedback (Price et al., 2010) or using rubrics or marking sheets (Denton & Rowe, 2015).

2.1.1 Atomic feedback

In this research project, we take a slightly different approach to provide written feedback to handwritten mathematics tasks more efficiently. After all, handwritten tasks remain important to train higher-order thinking skills and genuine problem-solving in mathematics education as students can express themselves more freely (Bokhove & Drijvers, 2010; Hoogland & Tout, 2018). Therefore, we propose a semi-automated (SA) approach: handwritten solutions are scanned, then teachers write feedback items, and the computer saves them so they can easily be reused when other students make similar mistakes (Moons et al., 2022; see Chapter 1).

How to write feedback that can easily be reused for other students? Long pieces of classic feedback are often too targeted to a specific student (Winstone et al., 2017). Hence, we suggest atomic feedback (see Figure 2.1): a collection of form requirements for written feedback that have been shown to make feedback significantly more reusable (Moons et al., 2022; see Chapter 1). To write an atomic feedback item, teachers must:

1. identify independent errors,
2. write small feedback items for each error separately, or
3. if an error reflects a structural mistake/misconception (Gusukuma et al., 2018), create two feedback items:
 - (a) one item containing feedback on the misconception in general and
 - (b) one or more sub-items addressing specific mistakes.

Atomic items ultimately form a point-by-point list covering only items relevant to a student’s solution. The list can be hierarchical in order to *cluster* items that belong together. Clustering ensures that feedback items can be written as atomically as possible. It prevents teachers from writing overly specific items because it provides an orderly way to present related feedback to students (e.g., through thematic clustering or a visual presentation of both general and specific feedback on the same error).

A comparison of classic (PP condition) and atomic feedback (SA condition) is presented in Figure 2.1. This comprehensive example demonstrates that classic feedback reports can be rephrased as atomic. This chapter compares all feedback reports from the PP and

Student's solution
Manipulate the formula: $A = 2\pi r h + 2\pi r^2$ to h

$$\frac{A}{2 \cdot \pi \cdot r} = h + 2\pi r^2$$

$$\frac{A - 2\pi r^2}{2 \cdot \pi \cdot r} = h$$

Classic feedback

Mind the fact that the dominant operation on the right-hand side of the equation is an addition! The division of the left-hand side by $2\pi r$ is, therefore, not helpful. Moreover, $2\pi r$ is a common factor of the right-hand side, but the sum wasn't completely divided by it (second addend not divided). Although your final answer is correct, the way it is written makes it look like a coincidence. Going from the first to the second step, you would normally subtract $2\pi r^2$ from both sides, meaning that it shouldn't be placed directly in the numerator, as you should make the denominators the same.

Atomic feedback

- First step
 - Dominant operation on the right side is an addition!
 - * Division of left-hand side is not helpful
 - * $2\pi r$ is a common factor of the right side, but:
 - sum wasn't completely divided by it
 - the second addend was not divided
- Second-step
 - Your final answer is correct, but:
 - * It looks like a coincidence.
 - * You should subtract $2\pi r^2$ from both sides.
 - * Mistake with making the denominators the same!
 - $2\pi r^2$ shouldn't be directly in the numerator.

Figure 2.1 – A comparison of classic and atomic feedback

SA conditions. In the SA condition, teachers were encouraged to write atomic feedback, but it is important to mention that all SA feedback reports will be considered; and not all of them adhere to the definition of atomic feedback (see [Chapter 1](#))¹.

2.1.2 Research aims

In Moons et al. (2022; see [Chapter 1](#)), it was demonstrated that feedback items meeting the atomic feedback requirements were significantly more reused than non-atomic items ($p < .001$, odds ratio: 2.6). This finding suggests that writing feedback items atomically enhances their reusability. Additionally, no significant differences in time investment were observed between the PP and SA conditions. However, teachers participating in the SA condition wrote significantly more feedback characters compared to the PP condition ($p = .02$, Cohen's $d = 0.41$), approaching a medium effect size. Despite these findings, an important research question remains unanswered:

[RQ 2.1] What similarities and differences do the SA and PP feedback types have regarding form characteristics, content characteristics and quality?

¹An important difference from the previous chapter is that this chapter considers feedback *reports*: the whole feedback text given to a student's solution to a question. In [Chapter 1](#), we categorised feedback *items* from the SA condition for being atomic or not. These *items* were the separate list entries and not the full reports.

To address this question, we will employ text mining techniques (Ferreira-Mello et al., 2019) and conduct a qualitative analysis (MacLure, 2013) on the feedback from both conditions. The qualitative analysis will investigate content characteristics and quality by coding the feedback reports. Through text mining, we will analyse word frequencies, sentiment, bigrams, and word correlations to compare the form and content characteristics of the two feedback types.

By addressing this research question, we aim to achieve two broader objectives. Firstly, we seek to gain a deeper understanding of how the utilisation of a statement bank, specifically reusing feedback, influences the characteristics of the resulting written feedback. This investigation will shed light on the impact of utilising pre-existing feedback statements on the form and content of the feedback provided. Secondly, we aim to explore the methodological approach of combining text mining and qualitative analysis to compare feedback. While text mining has been extensively used in higher education to analyse student course feedback (e.g., Grönberg et al., 2021), and qualitative approaches have been employed in combination (Hujala et al., 2020); the integration of these methodologies to compare feedback represents a relatively novel and promising application.

2.2 MATERIALS & METHODS

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

2.2.1 Materials

2.2.1.1 Semi-automated assessment tool for SA and text box for PP

For the SA condition, a self-developed plug-in in Moodle was used. While providing feedback on students' solutions, teachers always had three options in this condition: formulating atomic feedback, indicating that a solution was perfect, or indicating that the question was not answered (Figure 2.2a). They were able to use keyboard shortcuts to create a hierarchical list of feedback items. When a teacher typed something, the system searched the feedback items that had already been entered to detect possible matches for auto-completion (Figure 2.2a). The system searched only within the feedback items that the teacher had already entered for that particular question. In the PP condition, the teachers received only a text box to type feedback (Figure 2.2b), with no possibility of reusing feedback. In both conditions, teachers were also asked to give each solution a score out of 10.

2.2.1.2 Test on linear equations

We developed a test on linear equations in cooperation with a ninth-grade math teacher for this study. The test consisted of three items: 1. solving an equation, 2. manipulating a formula (see Figures 2.1 and 2.2), and 3. a modelling question consisting of a word problem (see Figure 2.8). The three items were combined to form a traditional test on linear equations. Solutions of 60 ninth-grade students (14-15 years old) from one

2. Manipulate the formula to h

$$\frac{A}{2 \cdot \pi \cdot r} = r + 2 \pi r^2$$

$$\frac{A - 2 \pi r^2}{2 \cdot \pi \cdot r} = h$$

Atomic feedback	Perfect (Max)	Not answered (0)
-----------------	---------------	------------------

- First step
 - Dominant operation on the right-hand side is an addition!
 - Division of the left-hand side is not helpful
 - $2\pi r$ is a common factor of the right-hand-side, but:
 - sum wasn't completely divided by it
 - second
 - the second addend was not divided

Calculated grade: 2/10

2 /10

2. Manipulate the formula to h

$$\frac{A}{2 \cdot \pi \cdot r} = r + 2 \pi r^2$$

$$\frac{A - 2 \pi r^2}{2 \cdot \pi \cdot r} = h$$

Your feedback:

Mind the fact that the dominant operation on the right-hand side of the equation is an addition! The division of the left-hand side by $2\pi r$ is, therefore, not helpful. Moreover, $2\pi r$ is a common factor of the right-hand side, but the sum wasn't completely divided by it (second addend not divided). Although your final answer is correct, the way it is written makes it look like a coincidence. Going from the first to the second step, you would normally subtract $2\pi r^2$ from both sides, meaning that it

2 /10

Figure 2.2 – Screens of the tool in the SA condition (a) & PP condition (b)

secondary school in Flanders (Belgium) were used in this study. The test and solution key can be found in [Appendix A](#).

2.2.2 Methods

2.2.2.1 Teacher participants

45 secondary mathematics teachers from Flanders with at least 3 years of working experience volunteered to participate in the study (28 female, 17 male). They were sampled using announcements in math teaching magazines. The average age of the participating teachers was 40.2 years ($SD = 10.3$). The first time we organised the experiment with 9 math teachers, we noticed several methodological imperfections in our study design. Therefore, this first attempt was used as a pilot study ($n = 9$) to refine the actual study ($n = 36$). A description of the pilot study and the adaptations to the actual study can be found in [Appendix B](#).

2.2.2.2 Study design

The study was set up as a crossover study (Bose & Dey, 2009) with two conditions: SA and PP. During a full working day, the teachers started in one condition in the morning and swapped to the other condition in the afternoon. The experiment was executed in the summer of 2020. Unfortunately, due to COVID-19 measures, only 9 participants at a time were allowed in the computer lab. Therefore, the experiment was repeated 7 times.

Each teacher gave feedback to all 60 solutions of the linear equation test, with a quasi-random selection of 30 solutions being assessed under the SA condition and the other 30 solutions under the PP condition. To mitigate order effects inherent in crossover experiments (Ratkowsky et al., 1993), half of the teachers started under the SA condition, and the other half started under the PP condition. The day started with training for all participating teachers. The training focused on working with the SA tool and the PP-text

field in Moodle and on how to formulate atomic feedback. The linear equation test was never mentioned during the training, and a geometry task obtained from other students was used instead as a demonstration. At the end of the training, teachers were asked to treat the students' solutions in the experiment in the same way that they would treat their own students. No training was provided in providing content-rich feedback and they were not informed about the research questions. Teachers had to be themselves above all.

The quasi-random selection of 30 solutions in each condition for each teacher ensured: (1) Comparability of the feedback between the conditions. Each solution was included in the SA condition of 18 teachers and the PP condition of the other 18 teachers, ensuring that both conditions comprised of feedback to the same solutions an equal number of times. (2) To balance the conditions for each teacher, we included 10 good, 10 moderate, and 10 bad students' in each condition based on the grades provided by the teachers of the pre-study. (3) The order in which the solutions were presented in each condition was random to avoid any bias caused by task familiarity or fatigue (Lim et al., 1996).

2.2.2.3 Data analysis procedures

Text mining

First, the provided feedback was explored using text mining techniques (Kwartler, 2017; Silge & Robinson, 2017). Text mining transforms unstructured text into a structured format to identify meaningful patterns and new insights using computer algorithms. It can be seen as a qualitative research method 'using quantitative techniques' (Yu et al., 2011).

A difficulty in applying text mining techniques is that many possible analyses can be employed. As this chapter aims to compare the given feedback to the same mathematics tasks in two conditions, we carefully applied techniques allowing us to find differences and similarities between these two feedback types. More specifically, we compared word frequencies, did a sentiment analysis, compared the Markov chains of bigrams and the pairwise correlations for both feedback types. These techniques were inspired by the book of Silge and Robinson (2017). More advanced approaches, such as LDA topic modelling, were executed but did not provide meaningful insights for our research question and are, therefore, not reported. We deliberately left out any significance tests in the text mining part as these tests are often over-powered when analysing on the level of words, making the sample sizes too large (Faber & Fonseca, 2014) or the test is executed on outcomes of an analysis that requires cautious interpretation (such as sentiment analysis), further supporting our decision. All the analyses were done using R.

Since the teachers participating in the study provided feedback in Dutch, all analyses were conducted in this language. In the pre-processing data phase, we removed all Dutch frequently used words (like 'a', 'the', 'of' in English) using a pre-defined lexicon (Benoit et al., 2021), a conventional first step in text mining analyses. In the final data analysis step, the results were automatically translated to English using the DeepLr-package (Zumbach & Bauer, 2021) to make the results interpretable for an international audience; hereby losing some specific language characteristics of Dutch (abbreviations, concatenations).

Qualitative analysis

For the qualitative exploration of the feedback, Busch et al. (2015a, 2015b) developed an instrument to assess teachers' diagnostic competencies. The categorical assessment tool of Busch et al. (2015a) was adapted to the particular research context described below.

Two authors of this chapter acted as independent raters for the qualitative analysis who coded blindly, meaning the coders could not see each other's codes in the process. The level of analysis was the full feedback report of the teacher on a student's solution to the word problem (see Figure 2.8). Four iterations were necessary to arrive at a high inter-rater reliability; the final Cohen's kappa coefficients can be found in Table 2.2. A random selection of about 100 feedback reports from the pilot study was used in each iteration. At the end of every iteration, the differences in coding were thoroughly discussed and some reconsidered. After establishing the final codebook, the coding process started on the actual study data. Feedback reports were coded student by student, and all feedback was checked for correctness by placing the student's solution next to it. Each researcher coded the teachers' feedback from 30 of the 60 students.

The codebook consists of **categorisable feedback** and **not further categorisable feedback**. The latter consists of **erroneous feedback**, **incomprehensible feedback** or **only addressing a solution was perfect, totally wrong or left blank**. A feedback report can only have one of these codes, meaning that erroneous or incomprehensible feedback reports were deliberately excluded from further classification.

The remaining categorisable feedback is to be coded in 5 sub-categories:

Concreteness judges how 'specific' the feedback is. For example, feedback containing only 'Order of operations!' is **general**, while ' $x = 14.4$? This can not be the number of answers!' points to **concrete** feedback. As a guideline, the two independent raters used the following question to decide between general and concrete: can this feedback without any adjustment be applied to another student who did something else? If yes, the feedback is classified as general; if not, the feedback is concrete. Concrete and general were mutually exclusive: as soon as something concrete was mentioned in the feedback, the whole report was characterised as concrete.

The **focus of the diagnosis** counts how many **deficits** and **strengths** the feedback addresses.

The **diagnostic activity** differentiates between **analysis**, **description** or **correction**: correction entails a teacher pointing to a mistake and giving the right solution (e.g., 'amount of correct answers: $30-4-x$ '). In contrast, description references a teacher addressing deficits without correction (e.g., 'wrong equation'). Finally, analysis signifies the teacher interpreting the student's mistakes and reporting that interpretation as feedback (e.g., 'You swapped answer and number of correct points in the choice of the unknown, $5x$ is the number of points gained with the correct answers, x the number of correct answers'). Merely noticing an error is seen as description, merely giving the right solution as correction. To ensure inter-rater reliability, a feedback report could only be coded into one diagnostic activity. When several diagnostic activities were identified in a feedback report, analysis was always preferred over correction, and correction always over description.

The **quality features of the diagnosis** contain four aspects, which were not mutually exclusive:

- **Explanation for deficits available:** the feedback contains a statement explaining why something is wrong in the solution (e.g., *'Why subtraction? Points should be added!'*). Explanation as a quality feature should not be confused with the diagnostic activity analysis: it can also be a more general expression of a mistake without interpretation at the student level.
- **Gives hints for improvement:** the feedback contains statements indicating how the solution should be improved in a possible future review (e.g., *'Keep points and number of questions well apart!'*). A hint can not contain the correct solution since the need for a future overhaul is then eliminated.
- **Notes that parts are missing in the student's solution:** the feedback explicitly refers to something that should have been in the solution (e.g., *'Write down the choice of the unknown'*).
- **Points to misconceptions:** the feedback contains statements to known misconceptions in mathematics education (Movshovitz-Hadar et al., 1987) or misunderstandings in the student's reasoning (e.g., *'You fail to see that your solution is impossible since you obtain more correct answers than there are questions.'*).

2.3 RESULTS & DISCUSSION

2.3.1 Text mining

2.3.1.1 Comparing word frequencies

A common first step in text mining is to compare word frequencies (Silge & Robinson, 2017). The frequency of a word is the proportion of the number of times a word occurs out of the total word count. Figure 2.3 gives a scatter plot of the used words in both feedback conditions. Words close to the identity line have similar relative frequencies in both conditions. It is apparent from this plot that most words scatter around this line, meaning that the majority of the words appear in both feedback types with a similar relative frequency. For example, 'attention' and 'both' appeared almost equally frequently in SA and PP. The observation that most words appeared in both feedback types with an almost equal relative frequency was confirmed by calculating Pearson's correlation coefficient of the word frequencies in both feedback types. It returned a high, positive correlation of $r(928) = 0.89$ with 95% CI [0.87, 0.90].

Words far from the identity line are, proportionally speaking, found more in one feedback type than the other. For example, 'super' and 'beautiful' were found more in PP feedback, while 'perfect' was found more in SA feedback. A likely reason is the default presence of a 'Perfect'-button that could be used for correct solutions in the SA condition. In the PP condition, teachers always had to write something themselves, and it seems they naturally chose a more diverse range of encouraging words. Also notable is the increased presence of many abbreviations in the PP condition, which DeepL understandably failed to translate, like 'opl' (Dutch abbreviation for 'solution'), 'vd' (= 'of the'), 'ptn' (= 'points') or 'antw' (= 'answer'). Teachers shortening feedback is one of the well-known coping

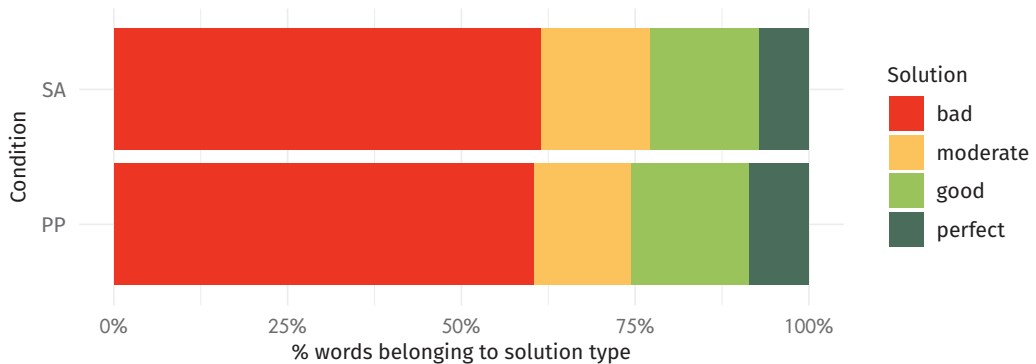


Figure 2.4 – Comparing the distribution of words spent on different solution types

The distribution of both feedback types looks essentially the same: proportionally, an almost equal amount of words is spent on bad solutions. SA feedback features slightly more feedback on moderate answers than PP feedback, which has proportionally more words coupled to good and perfect answers.

Although the word distribution in Figure 4 is some kind of sentiment analysis, in text mining, analysing the sentiment of a text is often done by using a pre-existing lexicon that assigns a polarity score to individual words (like ‘beautiful’ = 1, ‘incorrect’ = -1); subsequently, the sentiment of the whole text can be determined by taking the mean (Silge & Robinson, 2017). For the Dutch language, the Pattern lexicon (De Smedt & Daelemans, 2012) gives words a polarity score ranging from -1 (very negative), 0 (neutral), to 1 (very positive). For example, the following PP feedback has a mean polarity score of -0.65 (negative to very negative):

“Wrong choice of the unknown. A solution is found by guessing. However, guess cannot be right, you cannot give 95 wrong answers to 30 questions. No check.”

In contrast, the SA feedback below received a mean polarity score of 0.72 (positive to very positive):

- *Good choice of the unknown*
- *Good representation of the second unknown*
- *The equation that you have set up is perfect.*
- *The solution of the equation is perfect.*
- *You did not formulate an answer.”*

All feedback reports were analysed with this lexicon by taking the mean polarity score of all the words in the report. Next, we looked at the overall mean, the overall mean without the perfect solution type, and the mean for each solution type. The results can be found in [Table 2.1](#).

	Overall	Overall without perfect	Solution type			
			Bad	Moderate	Good	Perfect
SA	0.38 ± 0.60	-0.06 ± 0.41	-0.10 ± 0.38	-0.08 ± 0.39	0.10 ± 0.44	0.94 ± 0.20
PP	0.24 ± 0.45	0.02 ± 0.40	-0.01 ± 0.37	-0.04 ± 0.35	0.16 ± 0.47	0.62 ± 0.25

Table 2.1 – Mean polarity score and standard deviation overall and for each solution type

The sentiment analysis suggests that overall, the feedback in the SA and PP conditions has a neutral tone when perfect solutions are not considered. Moreover, the feedback tones are relatively equal when comparing the solution types in both conditions.

Lastly, some caution is necessary when interpreting this sentiment analysis. For example, the reason why we considered ‘overall without perfect’ as a separate column in [Table 2.1](#) is because including the perfect solutions induces a bias in favour of positive tones in the SA condition, as the button ‘perfect’ yielded feedback just saying ‘perfect’, with a polarity score of 1. The greater variety of appreciation words in the PP condition can sometimes include words or abbreviations not included in the sentiment lexicon; as such, the polarity score is sometimes estimated to be somewhat lower than 1, while the feedback reports are equally positive for these perfect solutions. Moreover, like in many sentiment analyses, the context was not taken into account; making statements like ‘not good’ having a polarity score of 0.6 as the ‘not’ is not seen as a word that reverses the polarity score; note, however, that the word usage of both conditions is almost equal (see previous paragraph), so the bias due to not including context is probably almost the same in both conditions.

2.3.1.3 Cluster analysis: Markov chains of bigrams & Pairwise correlations

To increase the readability of the plots in this paragraph, we limit ourselves to the feedback given in question 2 of the linear equations task (see [Figures 2.1](#) and [2.2](#)) in both conditions. [Figure 2.5](#) depicts the Markov chains of SA feedback (blue) and PP feedback (red). It visualises the pairs of consecutive words (= bigrams). As a cut-off, we have chosen a minimum of 10 co-occurrences. Although it represents a directed graph, we have omitted the arrows to increase readability.

Apparently, SA feedback features a denser linking structure between consecutive words. However, as reusing feedback is the main characteristic of this feedback, this was expected as some pairs will have been reused frequently, while PP feedback contains slight variations of word pairs. Nevertheless, some similar clusters arise in both feedback conditions. For example, noticing that double arrows should be used between the different intermediate steps was a cluster in both types. Interestingly, in the SA condition the word ‘notation’ also appears in this cluster. Using titles as a way of clustering feedback is one of the characteristics of atomic feedback, of which ‘notation’ is a clear example. If we examine the other clusters, other structuring elements in SA are found: ‘calculation rules’, ‘step 1’,... which do not appear in PP. SA with atomic feedback seems to foster teachers to structure feedback using titles, a phenomenon that does not emerge in PP feedback.

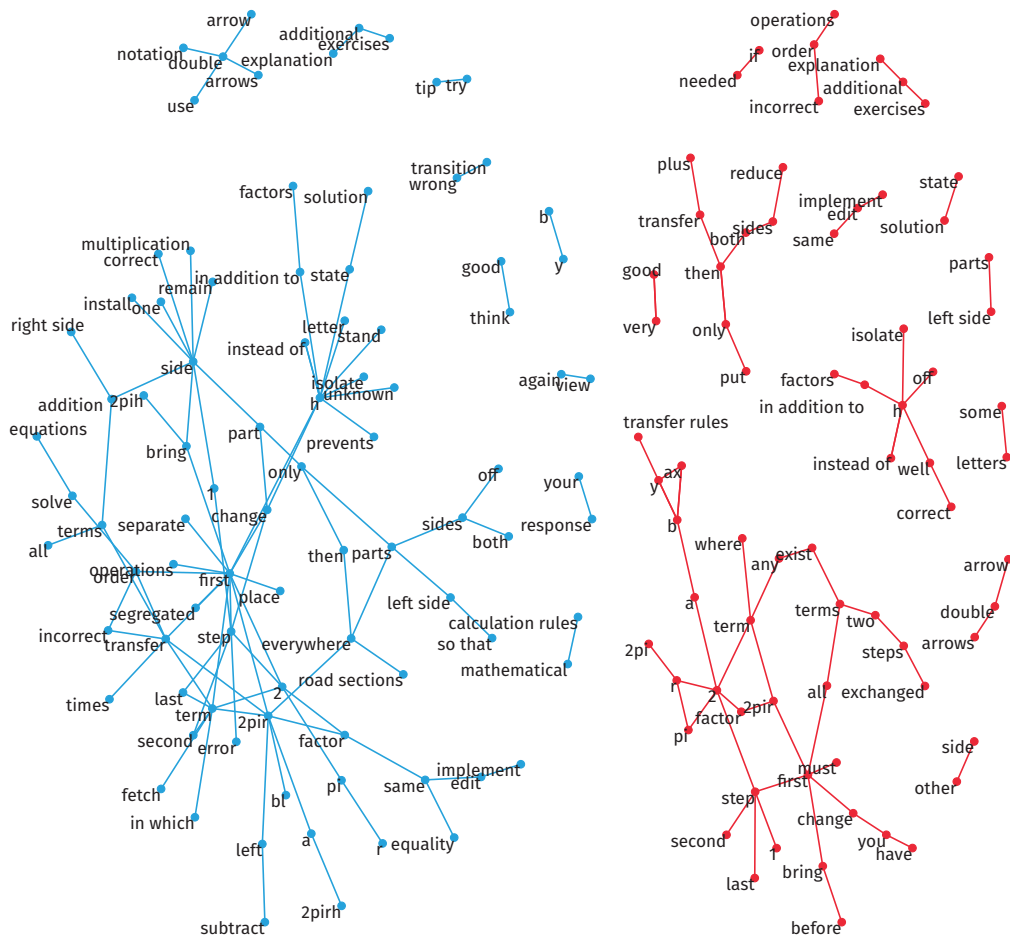


Figure 2.5 – Markov chains of bigrams for SA (blue) and PP (red) on question 2

Finally, pairwise correlations of the words in the same feedback reports were compared. Pairwise correlations differ from the bigrams in [Figure 2.5](#) as they do not link words succeeding each other but connect words often appearing together in the same feedback report (not necessarily consecutive). As SA contains many reused words, a denser correlation network is again to be expected. To marginally mitigate this bias in favour of SA, Spearman's rank correlation coefficients were used, comparing ranks instead of quantities. In [Figures 2.6](#) and [2.7](#), the correlation networks of the feedback on question 2 can be found. The different clusters refer to the same student's mistakes. Although the bias in favour of SA should be remembered, the difference in the largest cluster suggests that PP limits itself more often to short statements like 'reduce both sides' and 'isolate h '. In contrast, SA feedback seems to provide more information.

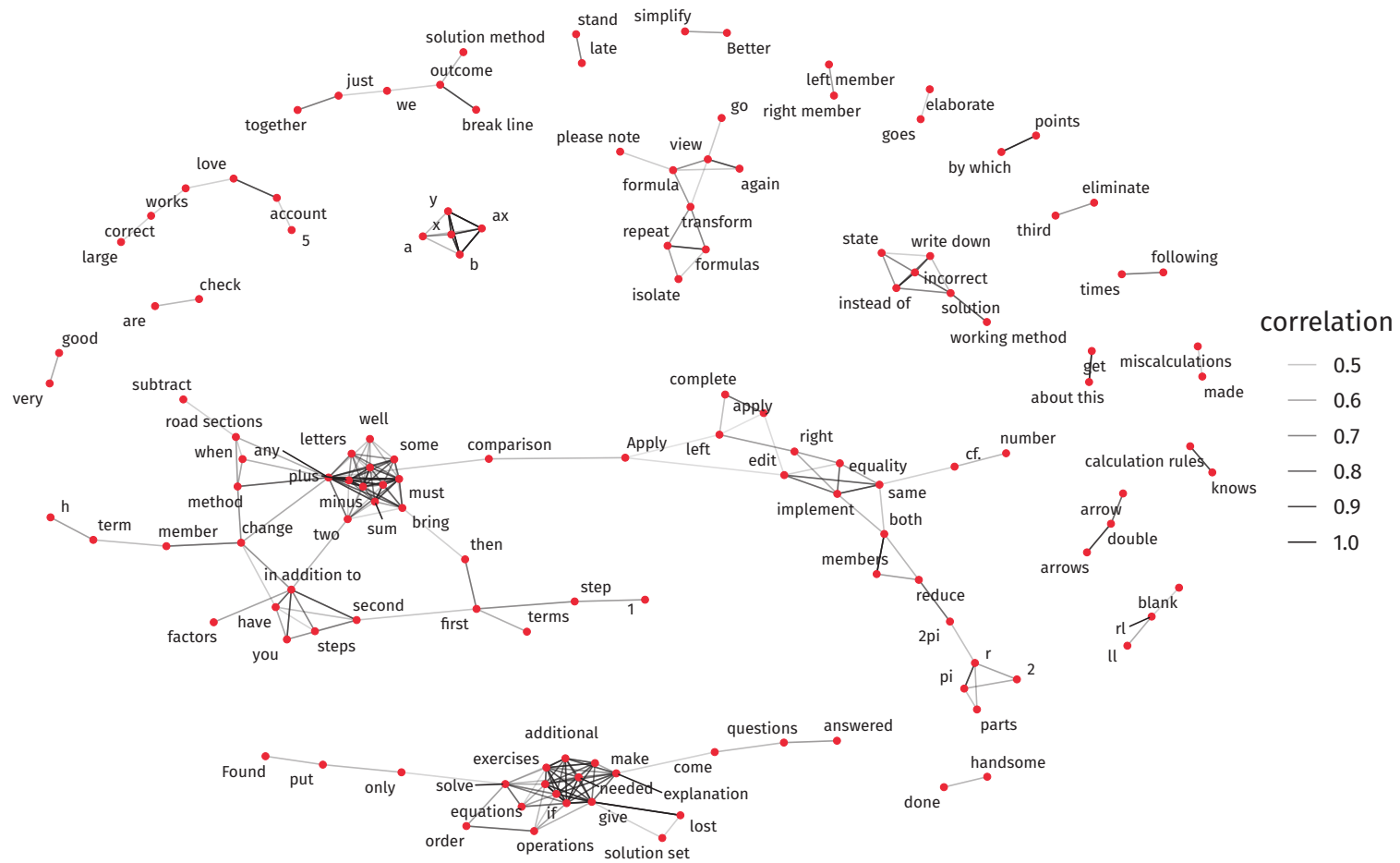


Figure 2.7 – Correlation network of PP feedback given to question 2

2.3.2 Qualitative analysis

Table 2.2 shows the results of the qualitative analysis. All percentages represent the proportion of feedback reports out of all feedback reports in that condition. The number of deficits and strengths between SA and PP were compared using a Mann-Whitney U test. All other reported p -values stem from two-sample z-tests for proportions, comparing for every category if the proportion of feedback items differs between SA and PP. The Pearson correlation coefficient ρ_{teacher} correlates the number of times a characteristic was chosen in both conditions for each teacher (or the number of deficits/strengths addressed). A strong ρ_{teacher} (> 0.7) for a characteristic indicates that the prevalence of the characteristic was consistent for teachers' feedback reports in both conditions. ρ_{student} reports the correlation on the level of the student solution.

	κ	SA ($n=913$)	PP ($n=947$)	p - value	ρ teacher	ρ student
Categorisable feedback		78.64%	78.04%	.631		
Concreteness						
General***	0.82	39.65%	30.52%	<.001	0.47	0.74
Concrete***	0.75	38.99%	47.52%	<.001	0.51	0.83
Focus of the feedback						
Number of deficits**	0.89 ¹	1.73 ± 1.28	1.57 ± 1.08	.003	0.48	0.87
Number of strengths*	0.89 ¹	0.79 ± 1.05	0.65 ± 0.84	.038	0.60	0.51
Diagnostic activity						
Analysis*	0.88	5.15%	7.60%	.038	0.22	0.51
Correction	0.87	16.21%	16.79%	.726	0.68	0.64
Description	0.84	56.63%	52.80%	.103	0.58	0.68
Quality features						
Explanation for deficits available*	0.58	9.42%	12.57%	.030	0.17	0.85
Gives hints for improvement	1.00	19.72%	23.23%	.072	0.68	0.68
Notes parts that are missing	0.49	15.55%	14.36%	.478	0.65	0.72
Points to misconceptions	0.82	4.93%	5.07%	.889	0.31	0.85
Not further categorisable feedback		21.36%	21.96%	.538		
Erroneous feedback	0.79	4.60%	4.96%	.711	-0.02	0.82
Incomprehensible feedback**	1.00	1.20%	0.11%	.003	-0.04	0.25
Only addresses						
...solution is entirely correct*	- ²	11.17%	14.68%	.024	0.19	0.68
...question is left blank*	- ²	3.40%	1.90%	.044	-0.20	0.90
...solution is entirely wrong	- ²	0.99%	0.32%	.072	0.52	0.30

* $p < .05$, ** $p < .01$, *** $p < .001$

¹ Intra-class correlation coefficient

² Automatically coded

Table 2.2 – Results of the qualitative analysis

2.3.2.1 Observed differences between SA and PP feedback

Overall, the results indicate SA feedback is less tailored to the student's solution than PP feedback: the SA reports are almost equally likely to be labelled as general or concrete (39.65% and 38.99%), whereas PP condition yielded much more concrete feedback (47.52%). However, SA seems more detailed: significantly more deficits and strengths were addressed in this condition; in contrast, PP seems more centred on the main issues in a solution. A frequently observed use of SA, which is general and can address different deficits and strengths, is using it as a sort of checklist, as the feedback below illustrates:

“■ *Choosing the unknown*

- *You are confusing the distinction between the number of questions and points received.*

■ *Setting up and solving the equation*

- *You did not include the unanswered questions*
- *Your equation is simpler than the equation to solve the question, but the solution is right.”*

Concerning the diagnostic activity, we see that there are significantly more feedback reports analysing where it went wrong with a student's solution in PP, from which an example is given:

“Please try again with x being the number of correct answers. Indeed, you know that for 26 questions, he got points. So you express the number of unanswered questions in terms of x . When setting up the equation, you noted 120 instead of 102. You have to take into account the 5 points per correct question.”

SA feedback reports tended to use more description and correction as diagnostic activity. Moreover, notice the low correlation $\rho_{\text{teacher}} = 0.22$ of teachers concerning analysis: teachers who analysed some solutions in one condition did not necessarily use that diagnostic activity as often in the other, suggesting that the SA system discourages teachers from providing feedback reports that analyse student's mistakes. One possible explanation is that teachers intuitively use SA too much as a checklist, preventing them from interpreting the interplay of intermediate steps the students took.

The significantly lower number of explanations given in the SA condition (the only significant difference in quality features) can be seen in the same vein. PP often addresses a particular mistake, on which the teacher sometimes gives an extra word of explanation. SA more often addresses all the mistakes in a solution, but treats these more superficially, without much extra information.

If we look at the differences in the ‘*not further categorisable feedback*’, we see that SA feedback is more often incomprehensible. However, it concerns only 1.2% of all feedback reports. Almost all of these stem from the same teacher who consistently used the hierarchical list of atomic feedback items in a confusing way by using opposite appreciation words in the parent items and child items, e.g.:

- “■ *Good formula*
- *Bad formula*”

It is striking that the readily available buttons ‘Perfect’ and ‘No answer’ in the SA condition had opposite effects. Just noticing a solution was ‘perfect’ or ‘cleverly done’ happened significantly more often without a button (!) in PP. This is consistent with the earlier observation that PP contains fewer deficits/strengths, but seems somewhat contradictory to the text mining analysis where ‘perfect’ was a more prevalent word in the SA condition. In the SA condition however, feedback reports for perfect solutions sometimes contained a complete list of all things that went well or noticed something that still could be added, such as a check if the obtained solution could be correct. In contrast, teachers did not hesitate to use the ‘No answer’-button in the SA condition when a solution was missing, while in the PP condition, they tended to give some hints how to start solving the question, wrote some encouraging statements, or asked the student what the underlying problem was (e.g., time issue, not enough understanding):

“The question was left blank. Did you have enough time?”

2.3.2.2 Observed feedback quality

The qualitative analysis allows us to evaluate the overall quality of both feedback conditions. A disappointing outcome that is shown in [Table 2.2](#) is that almost 1 out of 20 feedback reports is erroneous in both conditions. In other words, when feedback is handed out in an average classroom of compulsory education that, according to OECD (2012), consists of 21 students, one student will receive incorrect feedback. These errors might be due to routine, like teachers noting a common mistake that did not occur (they probably interpreted the student’s solution too quickly) or saying the solution is perfect, while the intermediate steps contain arithmetic errors. Nevertheless, more severe erroneous feedback was noticed, too: sometimes students choosing an alternative (but correct) solution path for the question and not arriving at the correct answer only received negative feedback in which their solution method was also (falsely) rejected. Luckily, erroneous feedback pops up coincidentally, as the within-teachers correlation of -0.02 shows it is not a consistent characteristic of teachers. However, some solutions lead to erroneous feedback more often in both conditions ($\rho_{\text{student}} = 0.82$).

The low proportion of *analysis* as diagnostic activity is also worrying. Part of the explanation is that some simple mistakes are not analysable such as a simple calculation error due to the absent-mindedness of the student: in such cases, a teacher can only notice the error. Consequently, the feedback would be coded as ‘description’ or ‘correction’. Nevertheless, analysis is not only lacking in these cases, but also when the student solution is well analysable like the one in [Figure 2.8](#). In this solution, the student makes a well-documented circular argument (Reusser & Stebler, 1997). By using the same given information twice, the student is left with an equation leading to an infinite number of solutions. Only 5 of the 36 teachers (14%) responded to this fallacy with feedback that analysed it; the other teachers gave descriptive feedback just noticing simple facts (e.g., ‘*equation is wrong*’) or corrective feedback. Of those five teachers, just one analysed this solution in the SA condition. SA feedback seems to engage teachers less in giving feedback analysing the student’s solution, compromising overall feedback

The Junior Mathematical Olympiad consists of 30 multiple-choice questions. You receive 5 points for each correct answer. Each wrong answer obvious results in 0 points, but you get 1 point for each empty question. In this way, Jurgen got a score of 102 points with 4 wrong answers. How many answers were correct?

1) ~~choosing~~ choosing the unknown

$$\begin{array}{l}
 x = \# \text{ correct answers} \\
 30 - 4 - x = \# \text{ unanswered questions} \\
 102 - 5x = \# \text{ points for } \begin{array}{l} \text{correct questions} \\ \text{unanswered} \end{array}
 \end{array}
 \left. \begin{array}{l} \\ \\ \end{array} \right\} \begin{array}{l} 102 - (102 - 5x) \\ = \# \text{ points for correct} \\ \text{answers} \end{array}$$

2) setting up and solving the equation

$$\begin{array}{r}
 102 = 102 - 5x + 102 - (102 - 5x) \\
 \Downarrow \\
 102 = 102 - 5x + 102 - 102 + 5x \\
 \Downarrow \\
 102 + 5x - 5x = 102 + 102 - 102 \\
 \Downarrow \\
 0x = 0 \quad \text{solution set: } \mathbb{R}
 \end{array}$$

3) answer you can not know how many correct answers he has because the solution set is \mathbb{R}

Figure 2.8 – An ‘analysable’ student’s solution to the word problem

quality. However, this example shows that factors other than the condition such as pedagogical content knowledge (Depaepe et al., 2013) or awareness of feedback quality criteria seem to play an important role for the feedback quality as well.

As mentioned in the introduction, shorter feedback is not necessarily worse for students (Chiles, 2021; Evans, 2013; Glover & Brown, 2006). However, the overall low number of deficits and strengths in Table 2.2 in both conditions gives pause for thought. While coding, we noticed a lot of ‘incomplete’ feedback reports, like the following PP feedback given to the solution in Figure 2.8:

“The first two lines are enough in your choice of the unknown. Equation is not set up correctly.”

One may wonder what students can learn from this feedback: they probably already figured out that the equation was incorrect, as infinite many correct answers seems a very unlikely outcome. And what about the other things they wrote? The phenomenon was seen many times while coding: feedback addressing the deficits at the start of the student’s solution; next, it concludes: ‘as a result, the rest of your solution is also wrong’, not saying anything about deficits and strengths in the rest of the student’s solution.

Some feedback seemed just too short to be meaningful to a student. This phenomenon occurred more in PP feedback as the number of addressed deficits is significantly lower.

2.4 CONCLUSIONS

To wrap up this chapter, we collected all our observations in [Table 2.3](#). With this explorative study comparing SA and PP feedback using text mining and qualitative techniques, we identified some essential characteristics of both feedback types.

First, we discovered similarities in both feedback types. From the text mining analysis, we distilled that the word usage and frequency is equal in both feedback types (S1), equal amounts of feedback were spent on bad, moderate and good solutions (S2), and feedback reports featured predominantly the same sentiments (S3). From the qualitative analysis, we know corrective and descriptive feedback appeared equally often as diagnostic activity in both conditions (S4), as well as giving hints, pointing at misconceptions and parts that are missing (S5). Writing erroneous feedback was also independent of the feedback type: it appeared almost equally often in both conditions (S6).

Many differences can be attributed to the observation of Price et al. (2010) that teachers often shorten feedback to reduce the workload of giving it. The need for this coping mechanism was profoundly reduced in the SA condition where teachers could reuse their feedback items: it contains more feedback (D1, Moons et al., 2022; see [Chapter 1](#)), fewer abbreviations (D2), addresses more mistakes and strengths (D5), and is more elaborate in describing mistakes (D9). However, this apparent comprehensiveness of SA feedback does not greatly improve the content quality: SA is often used as a checklist of all things that could go well/wrong, leading to more general feedback (D4). In contrast, PP feedback seems to be more focused on the main issues (D5), is more concrete and tailored to the student's solution (D4), and gives more short explanations of the observed deficits (D7, D9). More importantly, PP included more reports that analyse the student's solution (D6). When solutions were perfect, PP feedback used various appreciation words without much more, while SA often had some extra comments included in this case (D10); this is surprising as the SA condition featured a ready-to-use 'Perfect'-button, which was not present in the PP condition. In contrast, the ready-to-use 'No answer given'-button had the opposite effect (D8): in PP feedback, some encouragements or questions were included in the teachers' feedback when an question was left blank, while teachers in the SA condition mostly used the button. Finally, structuring elements like titles (D4) are an essential characteristic of SA; however, it is surprising that teachers did not naturally structure their feedback in the PP condition.

This study is not without its limitations. It is essential to acknowledge that we only tested the differences between PP and SA feedback on a mathematics task on linear equations in an experimental setting in which mathematics teachers gave feedback to the same 60 students' tasks and not on tasks of their own students. Next, the self-developed semi-automated assessment tool and the requirement for teachers to write atomic feedback make some similarities and differences rather specific for this research setting (e.g., the 'perfect' button), making not all of them applicable to the use of statement banks in general. Notwithstanding the study's limitations, we can still draw some lessons from this study. While our research question is answered in earlier

	SA feedback	PP feedback
Similarities	Stemming from text mining	
	(S1) Similar in both word usage and (relative) frequency	
	(S2) Equal distributions of feedback belonging to bad, moderate, and good solutions	
	(S3) Equal sentiments in both feedback types	
	Stemming from the qualitative analysis	
	(S4) Equal amounts of descriptive and corrective feedback	
Differences	(S5) Both give hints for improvement, note parts that are missing and point to misconceptions an almost equal amount of times	
	(S6) Almost 1 out of 20 feedback reports is erroneous	
	Stemming from a previous analysis	
	(D1) More feedback	(D1) Fewer feedback
	Stemming from text mining	
	(D2) Limited use of abbreviations	(D2) Abbreviations common
	(D3) Many structuring elements such as section titles	(D3) No structuring elements like titles
	Stemming from the qualitative analysis	
	(D4) More general, often used as a kind of checklist of right/wrong intermediate steps	(D4) More concrete and specific for the student's solution
	(D5) Addressess more deficits and strengths, including minor issues	(D5) Focuses mainly on main issues, less on minor deficits or strengths
	(D6) Feedback analysing the student's solution less common	(D6) Feedback analysing the solution more common
	(D7) Less explanations on mistakes	(D7) More explanations on mistakes
	Stemming from both text mining as the qualitative analysis	
	(D8) Empty questions get 'No answer' as feedback	(D8) Empty questions often receive an encouraging statement to get the student started
(D9) More elaborate feedback on mistakes	(D9) More short statements on mistakes	
(D10) Perfect solutions often labelled with 'perfect', although more often accompanied with side remarks	(D10) Perfect solution praised with a variety of appreciation words, with no extra remarks	

Table 2.3 – Observed similarities and differences between SA and PP feedback

paragraphs, we also aimed: (1) to learn how the use of statement banks (reusing of feedback) in general changes written feedback, and (2) how we can combine text mining with a qualitative analysis to compare feedback types. First, when statement banks are deployed carelessly, it naturally drags teachers into less effective ways of giving feedback compared to classic written feedback without statement bank. The feedback becomes more general, and the structure becomes centred on both major and minor aspects of the work that apply to many students, most likely because they can easily be repeated. Without a statement bank, the teachers' feedback is shorter but more focused on the main flaws in the students' work. Therefore, our main advice is to make teachers aware of this danger and that even when using statement banks, the rules of effective feedback remain key. Second, from combining text mining with 'classical' qualitative techniques, we learned that text mining gave us an overall idea about the differences and similarities in both feedback types; however, our qualitative analysis was essential to confirm some findings of the text-mining analysis. More importantly, we could only make statements on the content and quality of the feedback by using qualitative research methods. Text mining for education (Ferreira-Mello et al., 2019) is a promising research field, but in our view, not yet a self-sufficient methodology for comparing texts.

One critical follow-up question remains: how does a student interpret SA and PP feedback? We listed similarities and differences, but the litmus test is to see how students can act on the given feedback; a fruitful idea for a further research.

CRediT authorship contribution statement

Filip Moons: Conceptualisation, Methodology, Software, Formal analysis, Investigation, Writing – original Draft, Visualisation, Project administration, Resources, Funding acquisition. ● **Alexander Holvoet:** Formal analysis, Investigation, Writing – Review & Editing. ● **Katrin Klingbeil:** Conceptualisation, Resources, Writing – Review & Editing. ● **Ellen Vandervieren:** Supervision, Funding acquisition





PART II

**SEMI-AUTOMATED
ASSESSMENT FOR A GROUP
OF ASSESSORS**

HIGHLIGHTS

- ✓ A new chance-corrected inter-rater reliability measure is introduced allowing several raters to classify each subject into one-or-more categories. Some use cases include:
 - (1) Psychiatrists diagnosing patients into multiple disorders,
 - (2) Qualitative researchers categorising interview snippets into multiple codes of a codebook and
 - (3) Teachers choosing a couple of different criteria for student's work.
- ✓ The proposed statistic is a generalisation of Fleiss' kappa, which only allows several raters to classify each subject into one category. We can prove that when the data consists of one category for each subject, the proposed statistic reduces to Fleiss' kappa.
- ✓ The proposed κ statistic allows categories to be hierarchical and to have different weights of importance. It can handle missing data or a varying number of raters for each subject or category.
- ✓ The measure was discovered in order to answer the research question on blind versus visible checkbox grading in Chapter 4.



☑ Chapter 3

MEASURING AGREEMENT AMONG SEVERAL RATERS CLASSIFYING SUBJECTS INTO ONE-OR-MORE (HIERARCHICAL) NOMINAL CATEGORIES. A GENERALISATION OF FLEISS' KAPPA

PUBLICATIONS

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. (In preparation). Measuring agreement among several raters classifying subjects into one-or-more (hierarchical) nominal categories. A generalisation of Fleiss' kappa. *Psychological Methods*.

Moons, F. (2023). Measuring agreement among several raters classifying subjects into one-or-more (hierarchical) categories. A generalisation of the Fleiss' kappa [Master's thesis defended for the degree of MSc in Statistical Data Analysis, Ghent University]. (Grade: 18/20)

ABSTRACT

Cohen's and Fleiss' kappa are well-known measures for inter-rater reliability. However, they only allow a rater to select exactly one category for each subject. This is a severe limitation in some research contexts: for example, measuring the inter-rater reliability of a group of psychiatrists diagnosing patients into multiple disorders is impossible with these measures. This chapter proposes a generalisation of the Fleiss' kappa coefficient that lifts this limitation. Specifically, the proposed κ statistic measures inter-rater reliability between multiple raters classifying subjects into one-or-more nominal categories. These categories can be weighted according to their importance, and the measure can take into account the category hierarchy (e.g., categories consisting of subcategories that are only available when choosing the main category like a primary psychiatric disorder and sub-disorders; but much more complex dependencies between categories are possible as well). The proposed κ statistic can handle missing data and a varying number of raters for subjects or categories. The chapter briefly overviews existing methods allowing raters to classify subjects into multiple categories. Next, we derive our proposed measure step-by-step and prove that the proposed measure equals Fleiss' kappa when a fixed number of raters chose one category for each subject. The measure was developed to investigate the reliability of a new mathematics assessment method, of which an example is elaborated. The chapter concludes with the worked-out example of psychiatrists diagnosing patients into multiple disorders.

3.1 INTRODUCTION

Inter-rater reliability is the degree of agreement among independent observers who rate, code, or assess the same phenomenon. These ratings often rely on subjective evaluations provided by human raters, who sometimes differ greatly from one rater to another (Gwet, 2012; Vanacore & Pellegrino, 2022). Various researchers in many different scientific fields have recognised this problem for a long time since science requires measurements to be reproducible and accurate. Ideally, only a change in the subject's attribute should cause variation in the ratings, while the rater-induced source of variation should be excluded as it can jeopardize the integrity of scientific inquiries. The resolution to these problems, or at least the measurement of how big these problems are, is the study of inter-rater reliability.

The most well-known chance-corrected inter-rater reliability measures are Cohen's and Fleiss' kappa. However, these require *mutually exclusive categories*: a rater can only choose one category for each subject, it is not possible to classify subjects into multiple categories. Remarkably, very few attempts to lift this limitation are found in the literature. As such, the research question central in this chapter is:

[RQ 3.1] Can we develop a chance-corrected measure that allows multiple raters to classify subjects into one-or-more categories?

In the rest of this chapter, we briefly introduce Cohen's kappa, Fleiss' kappa, and their paradoxes. In 'other methods', we discuss the few attempts in the literature to lift the limitation of mutually exclusive categories. Next, we answer the research question by deriving the proposed measure. We start with the measure for regular categories, for which we can show that it is a generalisation of Fleiss' kappa. However, the measure can easily be extended to categories that differ in importance by giving them different weights and categories that exhibit a hierarchy of interdependencies. Finally, we compare the proposed measure with the described other methods from the literature by providing worked-out examples.

3.1.1 Cohen's kappa

Starting from the 1950s, various inter-rater reliability measures have been proposed (Bennett et al., 1954; Osgood, 1959), from which Cohen's kappa (1960) is the most well-known chance-corrected measure. This correction for chance is essential, as two raters may agree by following a clear, deterministic rating procedure, or they may agree by chance (Gwet, 2012). Thus, by accounting for chance, the kappa coefficient takes into account the difficulty of the classification task at hand. The formula of Cohen's kappa is:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \quad (3.1)$$

where P_o is the observed agreement and P_e is the expected agreement by chance. Cohen (1960) calls the numerator the *beyond-chance*: by subtracting the observed agreement with the expected agreement by chance, you are left with 'the percent of units in which beyond-chance occurred'; the denominator $1 - P_e$ can be seen as the 'beyond-chance' in the case of perfectly agreeing raters (the observed agreement is replaced with 1). So the kappa-statistic is the proportion of the observed beyond-chance over the beyond-chance in an ideal world of perfectly agreeing raters. Hence, the κ coefficient is the proportion of agreement *after* chance agreement is removed from consideration. κ coefficients always vary between -1 and 1 , with 1 indicating perfect agreement ($P_o = 1$), 0 indicating no agreement better than chance ($P_e = P_o$), and a value below zero indicates the agreement was less than one would expect by chance ($P_o < P_e$). The exact formulas for P_o and P_e for the Cohen's kappa can be found in Cohen (1960).

3.1.2 Fleiss' kappa

Cohen's kappa only allows to measure agreement between two independent raters, that is why Fleiss came up with the Fleiss' kappa in 1971 allowing a fixed number of 2 raters or more. These raters categorise subjects into exactly one of the available categories. We will now present how Fleiss defined P_o and P_e . Let I be the number of subjects, J is the (fixed) number of raters and C is the number of categories. Let x_{ic} be the number of raters who classified the i -th subject ($i \in \{1, \dots, I\}$) into the c -th category ($c \in \{1, \dots, C\}$). Since the categories are mutually exclusive, we know that every subject i will have received exactly J classifications, so $\sum_c x_{ic} = J$. We start with the observed agreement P_o . The extent of agreement among J raters for the subject i can be calculated as the proportion P_i of agreeing rater pairs $\binom{x_{ic}}{2}$ out of all the $\binom{J}{2}$ possible rater pairs. If x_{ic} equals 0 or 1, then there are no agreeing pairs, $\binom{x_{ic}}{2} = 0$. This proportion P_i for a subject

i can thus be defined as:

$$\begin{aligned} P_i &= \sum_c \frac{\binom{x_{ic}}{2}}{\binom{J}{2}} \\ &= \sum_c \frac{x_{ic}(x_{ic} - 1)}{J(J - 1)} \\ &= \frac{\sum_c x_{ic}^2 - J}{J(J - 1)}. \end{aligned}$$

The overall observed proportion of agreement P_o may then be measured by the mean of all P_i 's, so:

$$\begin{aligned} P_o &= \frac{1}{I} \sum_i P_i \\ &= \frac{\sum_i \sum_c x_{ic}^2 - IJ}{IJ(J - 1)}. \end{aligned} \quad (3.2)$$

We now turn to the formula of P_e , the expected agreement by chance. In total, IJ classifications will have been performed: all raters select exactly 1 category for each subject. So, the proportion of all assignments to the c -th category can be expressed as $\frac{\sum_i x_{ic}}{IJ}$, this is thus the probability to assign a subject to category c by chance. Consequently, the probability that any pair of (independent) raters classify a subject into category c by chance is given by $\left(\frac{\sum_i x_{ic}}{IJ}\right)^2$. Hence, if the raters made their classifications purely at random, the probability that two raters agree by chance on all categories is given by:

$$P_e = \sum_c \left(\frac{\sum_i x_{ic}}{IJ} \right)^2, \quad (3.3)$$

Plugging the above formulas into the κ statistic expressed in (3.1), gives the Fleiss' kappa:

$$\kappa = \frac{\frac{\sum_i \sum_c x_{ic}^2 - IJ}{IJ(J - 1)} - \sum_c \left(\frac{\sum_i x_{ic}}{IJ} \right)^2}{1 - \sum_c \left(\frac{\sum_i x_{ic}}{IJ} \right)^2}. \quad (3.4)$$

A more elaborate description and an example of psychiatric diagnosis on 30 subjects by six raters into a single disorder category, can be found in Fleiss (1971).

3.1.3 Paradoxes

Although both Cohen's kappa and Fleiss' kappa are widely popular measures for inter-rater reliability, some scholars have pointed out that these kappa coefficients are not free from paradoxes and can occasionally yield unexpected results (Feinstein & Cicchetti, 1990; Gwet, 2008; Warrens, 2010). One paradox arises when both the observed agreement P_o and the expected chance agreement P_e are high: the correction process embodied

in kappa's formula (3.1) can return a relatively low or even negative value of κ , whilst the observed agreement P_o is high. Another paradox is known as the prevalence paradox: it can be shown that the probabilities $\frac{\sum_i x_{ic}}{IJ}$ produce higher κ values when they are more balanced, i.e. when all categories are used about equally often and no particularly common categories exist. According to Gwet (2012), these probabilities are not suited to correctly measure the expected chance agreement P_e . All ratings for each category are used in the calculation of P_e , but as we want to say something about expected *chance* agreement, this philosophically implies we treat all these ratings as if they were all assigned randomly, which, according to Gwet (2012), is an unacceptable premise. Kraemer et al. (2002) disagree with Gwet's view, saying that 'it is well known that it is very difficult to achieve high reliability of any measure in a very homogeneous population (of subjects, ed.).'

3.1.4 Other methods

The literature on chance-corrected inter-rater reliability measures boomed in the 1970s and 1980s, with many proposals for different measures for different research settings. Surprisingly, only a few papers consider the limitation of mutually exclusive categories. This section briefly overviews the alternative methods in which a rater can classify a subject into multiple categories. Most of the methods below were described by Mezzich et al. (1981) but lacked sound mathematical expressions, which are added in this section.

3.1.4.1 Averaging or pooling Cohen's kappas

To calculate the inter-rater reliability among 2 raters who can classify subjects into multiple categories, a commonly used method is to calculate a Cohen's kappa for each category and average them: $\bar{\kappa}$ (De Vries et al., 2008). A problem with this approach is that when a category has an undefined Cohen's κ , $\bar{\kappa}$ is undefined too, which happens if the expected agreement by chance P_e is 1, e.g. when any rater did not select the category. A solution for this is *pooling* the Cohen's kappas by calculating the P_o and P_e for each category separately and then taking the average $\overline{P_o}$ and $\overline{P_e}$. Next, these averages are plugged in (3.1).

For example, NVivo (2022) - a popular program for qualitative research - advocates the pooled Cohen's kappa to measure the inter-rater reliability among two coders. These two coders (= 'raters') can code in NVivo the different sources (= 'subjects') of their research (e.g. text fragments, interviews, pictures) to one-or-more nodes of their codebook (= 'categories'). To get an overall κ of this coding process, Cohen's kappa is not suited: it would only allow the coders to code a source to exactly one node in their codebook. In contrast, a source is often coded to various nodes of the codebook. Therefore, Cohen's κ is calculated for each node in the codebook separately, and the pooled Cohen's kappa is used to get an overall κ of the coding process (see Figure 3.1). In 2008, De Vries et al. published a simulation study in which they compared 'true' Cohen's kappa values with the (simulated) averaged kappa and the (simulated) pooled kappa. Results showed that the pooled kappa almost always deviates less from the true kappa than the averaged kappa, resulting in smaller root-mean-square errors. Especially if the expected agreement by chance P_e is 0.6 or higher, the pooled Cohen's kappa outperforms the averaged Cohen's kappa. Indeed, when P_e is large, the denominator of the corresponding Cohen's kappa is small. In the case of the averaged kappa, the denominator of individual

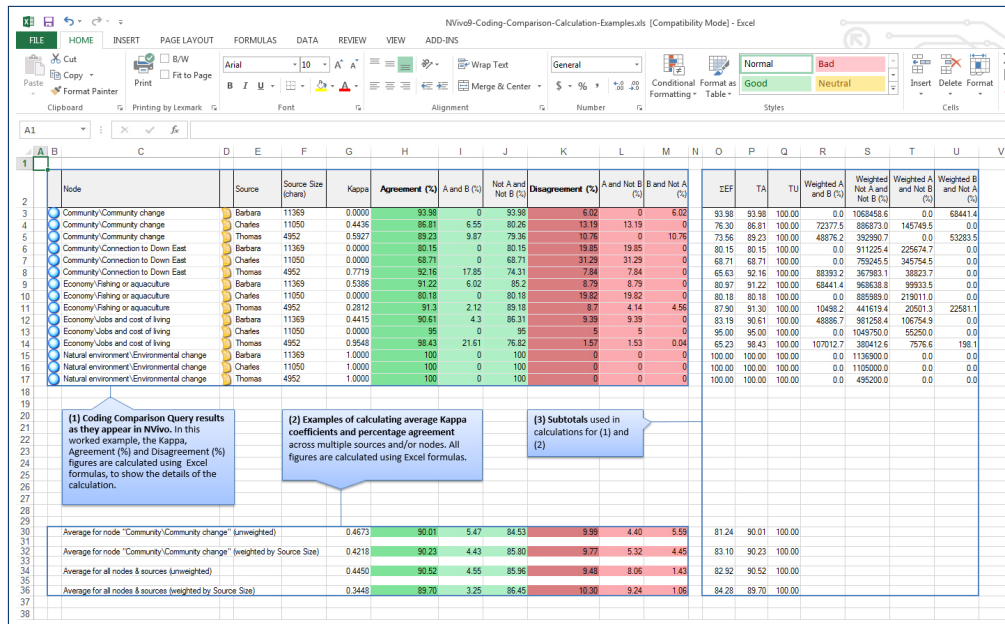


Figure 3.1 – NVivo advocates the pooled Cohen’s kappa approach in the provided Excel sheets to get an overall κ of the coding process (NVivo, 2022)

kappas has a multiplicative effect on the outcome (the numerator has an additive effect), making the method less precise when some individual denominators become small.

An important constraint to averaging or pooling Cohen’s kappas is embodied in the formulas of Cohen’s kappa itself: while the limitation of only one category for each subject is lifted, it is still limited to measure inter-rater reliability among exactly 2 raters. Moreover, it can not handle category hierarchies or different weights for categories.

3.1.4.2 Proportional overlap

The proportional overlap method was first introduced by Mezzich et al. in 1981. The method allows the calculation of a κ statistic in which multiple raters can classify subjects into multiple categories. The proportional overlap κ is calculated between pairs of raters. Let A_{ij} be the set of categories selected by rater j for subject i . The proportion of agreement between two raters a and b is then defined as the ratio of $\#(A_{ia} \cap A_{ib})$ (= the number of categories that were selected by both raters a and b for subject i) over $\#(A_{ia} \cup A_{ib})$ (= the total number of categories selected by rater a or b for subject i). For example, if rater a selected categories {blue, yellow, brown} and rater b selected {blue, green} for a given subject i , their proportional overlap is the ratio of 1 (one agreement on ‘blue’) over 4 (in total, rater a and b selected 4 different categories for subject i : blue, yellow, brown and green), so we get a proportional overlap of 0.25. In general, the proportional overlap ranges between 0 (= no overlap between the selected categories) and 1 (= perfect agreement, all categories match).

Agreement among several raters for a given subject is measured by averaging the proportional overlaps obtained for all combinations of pairs of raters for that subject. Next, the overall observed proportion of agreement P_o can be computed by averaging

these mean proportional overlaps for each of the I subjects¹:

$$P_o = \frac{\sum_i \sum_{(a,b) \in \binom{J}{2}} \frac{\#(A_{ia} \cap A_{ib})}{\#(A_{ia} \cup A_{ib})}}{I \frac{J(J-1)}{2}}$$

To determine the proportion of chance agreement P_e , we need to calculate the probability of two raters randomly agreeing on categories by chance alone. Mezzich et al. (1981) tries to achieve this by considering all possible combinations of two sets of selected categories A_{ij} (across all raters and subjects), computing their proportion of agreement, and taking the average of all these $\binom{I \cdot J}{2}$ proportional overlaps. This can easily be done using software and looping over all these combinations. It remains to be investigated whether the P_e definition of Mezzich et al. (1981) is entirely correct, as it is surprising it also considers combinations of selected categories A_{ij} of the same rater j , while these seem to be redundant in the calculation of P_e as the probability of two raters agreeing on categories should be considered. It is likely that the definition of Mezzich et al. (1981) slightly underestimates P_e (agreeing with oneself on random subjects will probably yield low proportional overlaps) and, as a result, slightly overestimates the κ statistic. A future simulation study can shed light on the matter. However, we stick for now the original definition of P_e and the mathematical rather complex formula is expressed below:

$$P_e = \frac{\sum_{(x,y) \in \binom{I \cdot J}{2}} \frac{\# [A_{\lceil x/J \rceil, (x-1 \bmod J)+1} \cap A_{\lceil y/J \rceil, (y-1 \bmod J)+1}]}{\# [A_{\lceil x/J \rceil, (x-1 \bmod J)+1} \cup A_{\lceil y/J \rceil, (y-1 \bmod J)+1}]}}{\frac{IJ(IJ-1)}{2}}$$

To understand the above formula, imagine an $I \times J$ -grid where each cell represents the classifications A_{ij} of subject i by rater j . This $I \times J$ -grid contains $I \cdot J$ cells that can be numbered row by row. An example with 3 subjects and 4 raters is given in Table 3.1. The last cell will have number $I \cdot J$. Now, take a pair (x, y) out of the possible combinations of two numbers from 1 to $I \cdot J$. Both x and y refer to a cell in the numbered $I \times J$ -grid. We need to translate x (and y) back to the corresponding subject and rater. To find the subject, take the ceiling of the division of x by J ($= \lceil x/J \rceil$). To find the rater, take $x - 1$ module J and add 1.

	Rater 1	Rater 2	Rater 3	Rater 4
Subject 1	$A_{11} \rightarrow 1$	$A_{12} \rightarrow 2$	$A_{13} \rightarrow 3$	$A_{14} \rightarrow 4$
Subject 2	$A_{21} \rightarrow 5$	$A_{22} \rightarrow 6$	$A_{23} \rightarrow 7$	$A_{24} \rightarrow 8$
Subject 3	$A_{31} \rightarrow 9$	$A_{32} \rightarrow 10$	$A_{33} \rightarrow 11$	$A_{34} \rightarrow 12$

Table 3.1 – Numbering an $I \times J$ -grid of classifications A_{ij}

As a number example with 3 subjects and 4 raters (Table 3.1), take the pair (10,12). To find the subject belonging to 10, we calculate $\lceil 10/4 \rceil = \lceil 2.5 \rceil = 3$, so we get subject 3. To

¹The notation $\sum_{(a,b) \in \binom{J}{2}}$ indicates a summation over all combinations of two raters. In this notation, $\binom{J}{2}$ denotes the set of all these combinations (and not the cardinality of this set).

find the rater belonging to 12, we calculate $(12 - 1) \bmod 4 = 11 \bmod 4 = 3$ and add 1, so we get rater 4.

The corresponding 'Mezzich's κ ' is found by plugging in P_o and P_e in Cohen's formula (3.1).

The proportional overlap method is an intuitive way to handle multiple raters classifying subjects into one or more categories and is easy to adapt to a varying number of raters (cf., some combinations of raters will not be present in this case). However, the method has limitations: it can not handle different weights for categories or category hierarchies. Moreover, the calculation of P_e depends on the number of combinations $\binom{I \cdot J}{2}$, which makes computation very demanding if the number of subjects I or the number of raters J is high. Using a random sample of combinations might solve the computational issue, but it is an open question of how large this random sample should be to guarantee sufficient accuracy.

3.1.4.3 Chance-corrected intraclass correlations

Mezzich et al. (1981) also proposed a method to use intraclass correlation coefficients as an intermediate step for the determination of a kappa statistic to allow the selection of multiple categories for each subject by multiple raters. To calculate the intraclass correlations, let $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \dots, x_{ijC})$ represent the classification vector of the i -th subject ($i \in \{1, \dots, I\}$) for the j -th rater ($j \in \{1, \dots, J\}$), with $x_{ijc} = 1$ when subject i was classified by rater j into category c ($c \in \{1, \dots, C\}$), and $x_{ijc} = 0$ otherwise. A measure of agreement is obtained by computing an intraclass correlation coefficient ρ_i between all \mathbf{x}_{ij} for a given subject i for all raters using a one-way ANOVA. If all the raters classified subject i in the same categories, perfect agreement is obtained, $\rho_i = 1$. P_o can be computed by taking the average of $\rho_1, \rho_2, \dots, \rho_I$. P_e is determined by computing the intraclass correlation coefficient between all classification vectors \mathbf{x}_{ij} for all raters and all subjects. Plugging P_o and P_e in (3.1) gives the value of the 'chance-corrected intraclass correlations.' Although the method is powerful by its simplicity, it can not handle different weights for categories, nor category hierarchies.

3.1.4.4 Chance-corrected rank correlations

The method proposed by Kraemer in 1980 is the only method we found in the literature where multiple raters classify subjects into an ordered list of categories: e.g., the best-fitting category for the subject according to the rater is ranked first, the second best-fitting category second, etc.

To calculate the corresponding kappa statistic, Kraemer uses classification vectors \mathbf{x}'_{ij} that contain ranks of the classifications drafted by rater j for subject i . For example, if for a given subject i , rater j made an ordered list of k categories ($k \leq C$), a 1 is assigned to the first category mentioned, a 2 to the second category, etc. Finally, categories that were not on the ordered list of rater j for subject i get rank $\frac{C+k+1}{2}$ assigned in vector \mathbf{x}'_{ij} , which equals the average of the remaining ranks. If raters can not decide the order between some selected categories, tied ranks can be placed in vector \mathbf{x}'_{ij} .

Assume, for example, that rater j made the following ordered classifications for subject i : 1. green 2. brown 2. orange 2. red 3. yellow., based on the 8 available categories {blue,

brown, green, pink, purple, orange, red, yellow} . Then, green would have a rank of 1, and brown, orange and red get rank 3 (i.e., the average of the ranks 2,3 and 4). Yellow receives rank 5. The unchosen categories (blue, pink, purple) get rank $\frac{8+5+1}{2} = 7$ (i.e., average of the remaining ranks 6,7 and 8). The resulting \mathbf{x}'_{ij} equals (7,3,1,7,7,3,3,5).

The chance-corrected rank correlations κ is calculated between pairs of raters. In this case, the Spearman correlation coefficient measures the agreement between two ranked classification vectors. Perfect agreement is obtained only if the two vectors are exactly the same. Let r_i be the average Spearman correlation coefficient between all pairs of raters for subject i , then P_o is the average of r_1, \dots, r_I :

$$P_o = \frac{\sum_i \sum_{(a,b) \in \binom{J}{2}} r_{\mathbf{x}_{ia}-\mathbf{x}_{ib}}}{I \binom{J-1}{2}}$$

P_e is calculated by averaging the Spearman correlation coefficient among all pairs of raters, for all subjects:

$$P_e = \frac{\sum_{(c,d) \in \binom{I \cdot J}{2}} r_{\mathbf{x}_{[c/J],(c-1 \bmod J)+1}-\mathbf{x}_{[d/J],(d-1 \bmod J)+1}}}{\frac{IJ(IJ-1)}{2}}$$

The corresponding κ is found by plugging in P_o and P_e in Cohen's formula (3.1). While the method is the only chance-corrected inter-rater reliability measure known in the literature allowing ranked classifications from raters, it can not handle different weights for categories nor category hierarchies. However, these probably do not appear in ranked classifications. The computational intensity for calculating P_e is the same as in the proportional overlap method.

3.2 DERIVATION OF THE PROPOSED KAPPA STATISTIC

3.2.1 Non-hierarchical categories

Suppose a sample of I subjects has been classified by the same set of J raters into C categories. The C categories are not mutually exclusive: a subject can be classified by a rater into multiple categories. Let $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \dots, x_{ijC})$ represent the classification vector of the i -th subject ($i \in \{1, \dots, I\}$) for the j -th rater ($j \in \{1, \dots, J\}$), with $x_{ijc} = 1$ when subject i was classified by rater j into category c ($c \in \{1, \dots, C\}$), and $x_{ijc} = 0$ otherwise. Let $x_{ic} = \sum_j x_{ijc}$ denote the number of raters classifying subject i into category c , with $J - x_{ic}$ representing the number of raters that did not classify subject i into category c . We can assemble all x_{ic} 's in an $I \times C$ -matrix X , containing all classifications. Some scholars would call X the 'agreement table.'

Furthermore, consider a weight vector $\mathbf{w} = (w_1, \dots, w_C)$ where w_c indicates the relative importance of category c proportional to the weights of the other categories. The choice of \mathbf{w} depends entirely on the research context in which the ratings took place. It is often conceptually convenient to impose $\sum_c w_c = 1$, but this is not required. In the unweighted case where all categories are equally important, we can take for all categories either $w_c = \frac{1}{C}$ or $w_c = 1$.

In case the categories are non-hierarchical, the selection of a category is independent from the (non-)selection of the other categories. The idea behind our proposed κ statistic is that we first derive a kappa statistic like the one described by Cohen (1960) for each category c :

$$\kappa_c = \frac{P_{O_c} - P_{e_c}}{1 - P_{e_c}}, \quad (3.5)$$

where P_{O_c} is the observed agreement for category c and P_{e_c} is the proportion of agreement expected by chance for category c . In our proposed κ statistic, the κ_c 's are not used directly, they solely give an impression on the agreement within each category separately. Instead, we will use the P_{O_c} 's and P_{e_c} 's and pool them together into one κ statistic.

We will calculate P_{O_c} pairwise (Conger, 1980). Two raters a and b agree on subject i when they both classified subject i into category c (so $x_{iac} = x_{ibc} = 1$) or when they both did not classify subject i into category c (so $x_{iac} = x_{ibc} = 0$). Hence, the extent of agreement for subject i and category c , can be seen as the proportion of rater pairs with agreement for category c to the total number of rater pairs. So, for subject i and category c , the numerator exists of the sum of $\binom{x_{ic}}{2}$ and $\binom{J-x_{ic}}{2}$, while the denominator is the amount of all possible rater pairs $\binom{J}{2}$. The proportion P_{ic} that denotes the extent of agreement for subject i and category c can thus be defined as:

$$\begin{aligned} P_{ic} &= \frac{\binom{x_{ic}}{2} + \binom{J-x_{ic}}{2}}{\binom{J}{2}} \\ &= \frac{(x_{ic})(x_{ic}-1) + (J-x_{ic})(J-x_{ic}-1)}{J(J-1)} \\ &= \frac{2x_{ic}^2 - 2Jx_{ic} + J^2 - J}{J(J-1)}. \end{aligned}$$

The overall observed proportion of agreement P_{O_c} for category c may then be measured by taking the mean of all P_{ic} 's so:

$$\begin{aligned} P_{O_c} &= \frac{1}{I} \sum_i P_{ic} \\ &= \sum_i \frac{2x_{ic}^2 - 2Jx_{ic} + J^2 - J}{IJ(J-1)}. \end{aligned} \quad (3.6)$$

P_{e_c} denotes the probability that two raters agree on (not) selecting category c by chance. For each category c , IJ decisions of (not) selecting c will have been performed. As x_{ic} denotes the number of raters classifying subject i into category c , $\sum_i x_{ic}$ represents the total number of classifications into category c . Hence, the proportion $\frac{\sum_i x_{ic}}{IJ}$ equals the probability that a rater randomly classifies a subject into category c . In case of two (independent) raters, the probability that both raters classify a subject into category c by chance is thus $\left(\frac{\sum_i x_{ic}}{IJ}\right)^2$. If x_{ic} raters classified subject i into category c , $J - x_{ic}$ raters did not. As such, the proportion $\frac{\sum_i (J-x_{ic})}{IJ}$ represents the probability that a rater did not

classify a subject into category c by chance. In case of two (independent) raters, the probability that both raters did not classify a subject into category c by chance is thus $\left(\frac{\sum_i (J - x_{ic})}{IJ}\right)^2$. Hence, the probability that two raters agree on (not) selecting category c by chance equals:

$$\begin{aligned}
 P_{e_c} &= \left(\frac{\sum_i x_{ic}}{IJ}\right)^2 + \left(\frac{\sum_i (J - x_{ic})}{IJ}\right)^2 \\
 &= \left(\frac{\sum_i x_{ic}}{IJ}\right)^2 + \left(\frac{IJ - \sum_i x_{ic}}{IJ}\right)^2 \\
 &= \left(\frac{\sum_i x_{ic}}{IJ}\right)^2 + \frac{I^2 J^2 - 2 \cdot IJ \cdot \sum_i x_{ic} + (\sum_i x_{ic})^2}{I^2 J^2} \\
 &= 2 \left(\frac{\sum_i x_{ic}}{IJ}\right)^2 - 2 \left(\frac{\sum_i x_{ic}}{IJ}\right) + 1
 \end{aligned} \tag{3.7}$$

We now aggregate all P_{o_1}, \dots, P_{o_C} and P_{e_1}, \dots, P_{e_C} into one kappa-statistic, including each category according to its weight w_c ²:

$$\kappa = \frac{\sum_c w_c (P_{o_c} - P_{e_c})}{\sum_c w_c (1 - P_{e_c})}. \tag{3.8}$$

If $\sum_c w_c = 1$ is imposed, this reduces to:

$$\kappa = \frac{\sum_c w_c P_{o_c} - \sum_c w_c P_{e_c}}{1 - \sum_c w_c P_{e_c}}.$$

3.2.1.1 The proposed κ statistic is a generalisation of Fleiss' kappa

When the requirements of the Fleiss' kappa are fulfilled, our proposed κ -static reduces to it:

Theorem

In case of equally weighted, mutually exclusive and non-hierarchical categories, the proposed kappa-statistic in (3.8) reduces to the Fleiss' kappa.

²Note that the proposed kappa statistic in (3.8) is not a weighted average of the individual k_c 's (3.5). In a yet-to-be-published simulation study, we can show that pooling the P_{o_c} 's and P_{e_c} 's in this way leads to smaller root-mean-square errors than using a weighted average of the k_c 's. This simulation study is similar to the study De Vries et al. (2008) did to compare averaging or pooling Cohen's kappa (see section 3.1.4.1). In addition to the smaller root-mean-square errors, this aggregation mechanism makes the κ statistic insensitive to undefined κ_c (e.g., when any rater did not select it, see the worked-out example in section 3.3.2).

Proof. As the categories are mutually exclusive, we know that $\sum_c x_{ijc} = 1$ for every combination of i and j , hence:

$$\sum_i \sum_c x_{ic} = \sum_i \sum_c \sum_j x_{ijc} = \sum_i \sum_j 1 = IJ. \quad (3.9)$$

Because the categories are equally weighted, we can take for all categories $w_c = \frac{1}{C}$, so we get:

$$\kappa = \frac{\sum_c \frac{1}{C} (Po_c - Pe_c)}{\sum_c \frac{1}{C} (1 - Pe_c)} = \frac{\sum_c (Po_c - Pe_c)}{\sum_c (1 - Pe_c)}.$$

First, we rewrite the denominator. Based on (3.7) and (3.9) we get that:

$$\begin{aligned} \sum_c (1 - Pe_c) &= \sum_c \left(1 - 2 \left(\frac{\sum_i x_{ic}}{IJ} \right)^2 + 2 \left(\frac{\sum_i x_{ic}}{IJ} \right) - 1 \right) \\ &= -2 \sum_c \left(\frac{\sum_i x_{ic}}{IJ} \right)^2 + 2 \sum_c \left(\frac{\sum_i x_{ic}}{IJ} \right) \\ &= -2 \sum_c \left(\frac{\sum_i x_{ic}}{IJ} \right)^2 + 2 \left(\frac{\sum_c \sum_i x_{ic}}{IJ} \right) \\ &= -2 \sum_c \left(\frac{\sum_i x_{ic}}{IJ} \right)^2 + 2. \end{aligned} \quad (3.10)$$

Second, based on (3.6) and (3.7), the numerator equals:

$$\begin{aligned} &\sum_c (Po_c - Pe_c) \\ &= \sum_c \left[\sum_i \frac{2x_{ic}^2 - 2Jx_{ic} + J^2 - J}{IJ(J-1)} - 2 \left(\frac{\sum_i x_{ic}}{IJ} \right)^2 + 2 \left(\frac{\sum_i x_{ic}}{IJ} \right) - 1 \right] \\ &= \sum_c \sum_i \frac{2x_{ic}^2 - 2Jx_{ic} + J^2 - J}{IJ(J-1)} - 2 \sum_c \left(\frac{\sum_i x_{ic}}{IJ} \right)^2 + 2 \sum_c \left(\frac{\sum_i x_{ic}}{IJ} \right) - C, \end{aligned}$$

applying (3.9):

$$\begin{aligned} &= \frac{2(\sum_i \sum_c x_{ic}^2) - 2JIJ + CIJ^2 - CIJ}{IJ(J-1)} - 2 \sum_c \left(\frac{\sum_i x_{ic}}{IJ} \right)^2 + 2 - C \\ &= \frac{2(\sum_i \sum_c x_{ic}^2) - 2IJ^2 + CIJ^2 - CIJ + 2IJ(J-1) - CIJ(J-1)}{IJ(J-1)} - 2 \sum_c \left(\frac{\sum_i x_{ic}}{IJ} \right)^2 \\ &= \frac{2(\sum_i \sum_c x_{ic}^2) - 2IJ}{IJ(J-1)} - 2 \sum_c \left(\frac{\sum_i x_{ic}}{IJ} \right)^2. \end{aligned} \quad (3.11)$$

Finally, we divide (3.11) by (3.10) and get the well-known Fleiss' kappa (3.4):

$$\kappa = \frac{\frac{(\sum_i \sum_c x_{ic}^2) - IJ}{IJ(J-1)} - \sum_c \left(\frac{\sum_i x_{ic}}{IJ} \right)^2}{1 - \sum_c \left(\frac{\sum_i x_{ic}}{IJ} \right)^2}.$$

□

Remark the apparent difference between P_{o_c} (3.6) and P_{e_c} (3.7) in the proposed measure and P_o (3.2) and P_e (3.3) in Fleiss' kappa: as Fleiss' kappa presumes mutually exclusive categories, two raters a and b can only agree on a subject i if they both classified the subject into the same category c , and there are $\binom{x_{ic}}{2}$ such agreeing rater pairs. Everything else can be regarded as a disagreement.³ This no longer holds when a subject i can be classified into multiple categories by the same rater: when raters a and b do not select category c for subject i , they agree that from all C categories that can be selected, category c should not be. So the number of agreeing pairs is the sum of $\binom{x_{ic}}{2}$ and $\binom{J-x_{ic}}{2}$; meaning that the agreement on not classifying subject i into category c , is valued equally as the agreement on an actual classification of subject i into category c by both raters a and b . This is a philosophical premise of this proposed κ statistic, and every user should consider whether this premise is appropriate in a specific context. If the proposed κ statistic is used with mutually exclusive, equally weighted, and non-hierarchical categories, the theorem in section 3.2.1.1 shows that all these terms of agreement on non-classification cancel out.

3.2.1.2 Handling missing data or a varying number of raters

Until now, we only considered the case of a fixed number of raters J . However, in practice, raters may only have classified a proportion of the participating subjects or even used only a proportion of the available categories. Two possibilities can be distinguished:

1. **Missing data:** some classifications of raters are lost due to unforeseen circumstances. However, the experiment was not designed not to collect this data.
2. **Varying number of raters:** raters only had the opportunity to rate a portion of the participating subjects or use only some of the categories. The experiment was intentionally designed to collect only this data (for example, for feasibility reasons).

To adapt our measure to handle both missing data and a varying number of raters, we need to take a step back and revisit our derivation of P_{o_c} . In section 3.2.1, we defined P_{o_c} as an average of the P_{i_c} 's over all subjects. This was done for didactic reasons, as it nicely shows the similarities with the construction of Fleiss' kappa (see section 3.1.2). However, P_{o_c} can also be seen as the proportion of rater pairs with agreement for category c to the total number of rater pairs. The number of rater pairs with agreement for category c can be calculated as a summation over all subjects: $\sum_i \binom{x_{ic}}{2} + \binom{J-x_{ic}}{2}$, the

³In fact, the calculation of κ_c (see (3.5)) is equal to the calculation of the Fleiss' kappa with two categories: 'selected category c ' and 'not-selected category c '.

total number of rater pairs too: $\sum_i \binom{J}{2}$, so we get:

$$\begin{aligned} P_{O_c} &= \frac{\sum_i \binom{x_{ic}}{2} + \binom{J-x_{ic}}{2}}{\sum_i \binom{J}{2}} \\ &= \frac{\sum_i [(x_{ic})(x_{ic}-1) + (J-x_{ic})(J-x_{ic}-1)]}{\sum_i J(J-1)} \\ &= \frac{\sum_i (2x_{ic}^2 - 2Jx_{ic} + J^2 - J)}{\sum_i J(J-1)}, \end{aligned}$$

as $\sum_i J(J-1) = IJ(J-1)$ and taking the sum of fractions with equal denominators results in the sum of the numerators divided by this denominator, it is clear that the formula above equals the formula in (3.6).

This other approach of deriving P_{O_c} is essential to adapt our proposed κ statistic (3.8) for handling both missing data and a varying number of raters. First, we replace the fixed number of raters J . Define the $I \times C$ -matrix J' , with the elements j_{ic} representing the number of raters that had the opportunity to classify subject i into category c . With the derivation above, adapting P_{O_c} is straightforward:

$$P_{O_c} = \frac{\sum_i \binom{x_{ic}}{2} + \binom{j_{ic}-x_{ic}}{2}}{\sum_i \binom{j_{ic}}{2}} = \frac{\sum_i (2x_{ic}^2 - 2j_{ic}x_{ic} + j_{ic}^2 - j_{ic})}{\sum_i j_{ic}(j_{ic}-1)},$$

for P_{e_c} (3.7), we get:

$$\begin{aligned} P_{e_c} &= \left(\frac{\sum_i x_{ic}}{\sum_i j_{ic}} \right)^2 + \left(\frac{\sum_i (j_{ic} - x_{ic})}{\sum_i j_{ic}} \right)^2 \\ &= 2 \left(\frac{\sum_i x_{ic}}{\sum_i j_{ic}} \right)^2 - 2 \left(\frac{\sum_i x_{ic}}{\sum_i j_{ic}} \right) + 1. \end{aligned}$$

In the case of missing data, these formulas imply the 'Missing Completely at Random' (MCAR)-assumption, as we estimate the values based on the available data and therefore see the available data as representative for the full data (Little, 1988).

Although the proposed κ statistic is flexible enough to handle missing classifications in some categories with the formulas above, this is often scientifically unacceptable: when raters do not have an overview of all categories, they will be forced to classify some subjects into different categories than they would have done with all categories available. Normally, only a varying number of raters for each subject is desirable. In that case, the matrix J' can be replaced by a vector $\mathbf{j} = (j_1, j_2, \dots, j_I)$ with j_i the number of raters who classified subject i , and P_{O_c} (3.6) and P_{e_c} (3.7) can be simplified accordingly:

$$P_{O_c} = \frac{\sum_i (2x_{ic}^2 - 2j_i x_{ic} + j_i^2 - j_i)}{\sum_i j_i (j_i - 1)}, \quad (3.12)$$

$$P_{e_c} = 2 \left(\frac{\sum_i x_{ic}}{\sum_i j_i} \right)^2 - 2 \left(\frac{\sum_i x_{ic}}{\sum_i j_i} \right) + 1 \quad (3.13)$$

3.2.2 Hierarchical categories

3.2.2.1 Actual classifications versus possible classifications

Let us now consider the case when categories have some kind of hierarchical structure. For example, the categories to which a rater classifies subjects can have main categories and subcategories; with a subcategory only be selectable if the main category was chosen. Also more complex hierarchical structures are possible: think of decision graphs in which some subcategories can only be chosen when some condition is met (e.g., a category can be selected when only one of two other categories is selected, a category can only be selected when another is not selected).

No matter how the hierarchical structure of the categories is constructed, all these hierarchies have one thing in common: based on the classifications rater j already made for subject i , some (sub)categories will (not) be selectable. In other words: where in the non-hierarchical case every subject i could be classified J times into category c , in the hierarchical case the upper limit of possible classifications of subject i into category c will depend on the number of raters who could select category c , we will denote these *possible classifications* as s_{ic} .

It is important to understand the difference between the s_{ic} 's and x_{ic} 's for a given subject i and category c : x_{ic} denotes the number of *actual* classifications of subject i into category c ; so the number of times category c was selected for subject i , while s_{ic} indicates the number of *possible* classifications of subject i into category c . This means that s_{ic} corresponds to the number of times category c was available for selection in case of subject i , which directly follows from the hierarchical structure of the categories. The calculation of s_{ic} for a given category c and subject i can depend on actual classifications of higher-order categories for subject i , but never on x_{ic} itself.

Let S be an $I \times C$ -matrix, with elements s_{ic} defined as the number of possible classifications of subject i into category c with $\forall i, \forall c : s_{ic} \in \{0, 1, \dots, J\}$ and x_{ic} can never exceed the number of possible classifications s_{ic} of subject i in category c , so $\forall i, \forall c : s_{ic} \geq x_{ic}$.

In the following section we will show that taking into account the hierarchy of the categories only depends on these s_{ic} 's to compute the κ statistic. To give an impression on how to calculate the s_{ic} 's: all main categories could be selected by all J raters for every subject i , so $s_{ic} = J$ for all main categories. In a simple parent-child hierarchical structure, a child category c' can only be selected if the parent category p was selected so $s_{ic'} = x_{ip}$, i.e. the number of *possible* classifications of child category c' for subject i equals the number of *actual* classifications in parent category p for subject i . For more complex hierarchical structures, the calculation of s_{ic} can depend on a couple of different x_{ijc} 's; apprehensive of the inclusion-exclusion principles of combinatorics (for an example, see the worked-out example in [section 3.3.1](#)).

3.2.2.2 The kappa-statistic

With the introduction of matrix S , the construction of Po_c and Pe_c is straightforward: replace every occurrence of J by the respective s_{ic} 's (3.6) and (3.7). Using the same

approach as section 3.2.1.2 to look at P_{O_c} , we get:

$$\begin{aligned}
 P_{O_c} &= \frac{\sum_i \binom{x_{ic}}{2} + \binom{s_{ic} - x_{ic}}{2}}{\sum_i \binom{s_{ic}}{2}} = \frac{\sum_i [(x_{ic})(x_{ic} - 1) + (s_{ic} - x_{ic})(s_{ic} - x_{ic} - 1)]}{\sum_i s_{ic}(s_{ic} - 1)} \\
 &= \frac{\sum_i (2x_{ic}^2 - 2s_{ic}x_{ic} + s_{ic}^2 - s_{ic})}{\sum_i s_{ic}(s_{ic} - 1)}, \tag{3.14}
 \end{aligned}$$

and for P_{e_c} :

$$P_{e_c} = \left(\frac{\sum_i x_{ic}}{\sum_i s_{ic}} \right)^2 + \left(\frac{\sum_i s_{ic} - x_{ic}}{\sum_i s_{ic}} \right)^2 = 2 \left(\frac{\sum_i x_{ic}}{\sum_i s_{ic}} \right)^2 - 2 \left(\frac{\sum_i x_{ic}}{\sum_i s_{ic}} \right) + 1. \tag{3.15}$$

If we would aggregate P_{O_c} and P_{e_c} in the same way as in (3.8), then we would have adjusted the contribution of category c according to the context-related weights w_c . However, in our aggregation, we would not have adjusted for the total possible classifications $\sum_i s_{ic}$ of category c . This is not desirable, which can be illustrated by the following example: suppose unweighted categories and assume that for a subject i only two raters could select subcategory c' , so $s_{ic'} = 2$. Rater 1 classified subject i into subcategory c' and rater 2 did not. Moreover, due to the category hierarchy, the subcategory c' was not available for all the other subjects for all raters, so $\sum_i s_{ic'} = 2$. This will lead to a $P_{O_c} = 0$ and $P_{e_c} = 0.5$. With no additional scaling for the total possible occurrences of a category (and thus using formula (3.8) for aggregating P_{O_c} and P_{e_c}), the subcategory will contribute -0.5 to the numerator and 0.5 to the denominator. In other words, if we do not adjust for possible classifications, we pull the value of κ down for an almost negligible category that was only a possible classification on two occasions. In contrast, the main categories had IJ possible classifications.

To solve the problem and adjust for the total possible classifications s_{ic} of category c , we introduce a scaling factor ϕ_c for each category c , to scale the terms $P_{O_c} - P_{e_c}$ in the numerator and the terms $1 - P_{e_c}$ in the denominator:

$$\phi_c = \frac{\sum_i s_{ic}}{IJ}. \tag{3.16}$$

This scaling factor contrasts the total possible occurrences of a category with the IJ possible classifications of main categories. As a result, main categories always have $\phi_c = 1$. With the expressions in (3.14), (3.15) and (3.16), we are now ready to define the kappa-statistic for the hierarchical case:

$$\kappa = \frac{\sum_c w_c \phi_c (P_{O_c} - P_{e_c})}{\sum_c w_c \phi_c (1 - P_{e_c})}. \tag{3.17}$$

3.2.2.3 Handling missing data or a varying number of raters

Note that in the calculation of the proposed kappa-statistic for hierarchical categories (3.17), only the scaling factors ϕ_c still refer to the assumption of a fixed number of raters J . A varying number of raters or missing data should therefore be handled within the

calculation of matrix S of possible classifications, with respect to the hierarchy of the categories. As in [section 3.2.1.2](#), we again introduce the $I \times C$ -matrix J' with the elements j_{ic} representing the number of raters that could have classified subject i into category c , *irrespective* of the hierarchy of the categories. This means that s_{ic} is only equal to j_{ic} in the case that c is a main category that is available under all circumstances to raters. In other words: j_{ic} represents the number of possible classifications of subject i into category c without prior knowledge of the other categories the raters have selected (in contrast, this knowledge is definitely required to calculate the matrix S). Hence, matrix J' is what we need to adjust the denominator of (3.16). The scaling factors ϕ_c adjusted for a varying number of raters are defined as:

$$\phi_c = \frac{\sum_i s_{ic}}{\sum_i j_{ic}},$$

If the number of raters only varies over subjects (and not over categories), matrix J' can be replaced by vector $\mathbf{j} = (j_1, j_2, \dots, j_I)$ with j_i defined as the number of raters who classified subject i ; the adapted κ statistic appears by changing matrix S and the scaling factors ϕ_c 's accordingly.

3.3 WORKED-OUT EXAMPLES

In this section, we apply our proposed κ statistic and the appropriate other methods from [section 3.1.4](#) to two applications: one on the assessment of a mathematics exam for which our proposed κ statistic was initially developed, the other is an example from Mezzich et al. (1981) in which 30 child psychiatrists diagnose patients into multiple psychiatric disorders.

3.3.1 Assessing mathematics exams

3.3.1.1 Context

The proposed κ statistic was initially developed to measure the inter-rater reliability of multiple assessors assessing students with a new assessment method (Moons and Vandervieren, 2022; see Chapters 4 and 5) for handwritten high-stakes mathematics exams called 'checkbox grading.' The method allows exam designers to preset a list of feedback items with partial scores for each question; so that assessors should just tick the items (= categories) relevant to a student's answer. Hierarchical dependencies between items can be set, so items can be shown, disabled, or adapted whenever a previous item is ticked, implying that assessors must follow the preset point-by-point feedback items from top to bottom. This adaptive grading approach resembles a flow chart that automatically determines the grade. Moreover, checking 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 designers and (2) a straightforward way to do correction work with multiple assessors where personal interpretations are avoided as much as possible.

An example of checkbox grading is given in [Figure 3.2](#). With this drawing question, a student can gain a maximum score of 3 points. If point A is drawn correctly (1st bullet),

Question (Max: 3 points)

Draw the line segment $[AB]$ with $co(A) = (-3, 4, 0)$ and $co(B) = (-3, 4, 5)$ in the given axis system.

- (1) Point A is drawn correctly +1.0
- (2) One or more auxiliary lines are drawn for point B
- (3) Point B is drawn correctly +1.5
Pay attention to the correct height!
- (4) Can only be assessed when A and B are correct.
- (5) Line segment AB is drawn correctly +0.5 if A and B are drawn correctly
- (6) Line AB is drawn -0.5 if A , B and $[AB]$ are drawn correctly

Solution key

Figure 3.2 – Example question of the mathematics assessment tool

the student gains 1 point; the correct drawing of point B (3rd bullet) is worth 1.5 points. The 2nd bullet does not change the score but shows assessors that the presence of auxiliary lines is perfectly fine. The last two feedback items, bullets 4 and 5, can only be selected if items 1 and 3 were selected. As the drawing of the line AB implies the drawing of the line segment $[AB]$, the 5th bullet can only be selected if the 4th was. This is a clear example of hierarchical items (= categories).

During the project, one of the main research questions concerned the inter-rater reliability of this new assessment method under two conditions: blind versus visible grading. As the computer automatically calculated the grade associated with the selected checkboxes, it was possible to hide the grades and calculation from the assessors, which was the blind condition. In the visible condition, assessors could see how the items influenced the grade and how the total grade was calculated. From the literature on rubrics (Dawson, 2017), we know that judges often change the selection of criteria when the resulting grade does not align with their holistic appreciation of the work, which can affect the instrument's reliability. As such, the research question was: 'Does blind checkbox grading enhance inter-rater reliability compared to visible checkbox grading?'

The traditional measures for inter-rater reliability such as intraclass correlations fell short because these can only measure the agreement between assessors on grades, while the method also provides feedback to students. Hence, it is not enough to agree on grades; the resulting feedback to the students must also be as equal as possible. Score agreement by no means guarantees agreement on feedback items, which is especially clear for feedback items not influencing the score (e.g., bullet 2 in the example). Other examples can be given as well: in Figure 3.2, 2.5 points can be obtained by solely drawing points A and B correctly (only bullets 1 and 3 apply, possibly bullet 2) or by drawing the line AB correctly (all bullets apply, possibly bullet 2). Conversely, the inverse is true: agreement on feedback items implies score agreement.

Our proposed κ statistic of section 3.2.2 does meet all requirements:

- It will assess the agreement of the raters in selecting multiple feedback items (= categories) for each student (= subjects)
- These items are hierarchical: the selectability of some items depends on the selection of other items
- Score agreement can naturally be measured by weighing the items according to their partial scores.

3.3.1.2 Example

We start with a worked-out example, in which our proposed κ statistic is calculated step-by-step. We consider 3 assessors (i.e., the number of raters J equals 3) assessing 6 students' solutions (i.e., the number of subjects I equals 6) on the question in Figure 3.2. The assessors classified every student's solution into the 5 checkboxes/feedback items (i.e., the number of categories C equals 5). The classifications by the three assessors of the six students' answers can be found in Table 3.2. Although the example consists of a simple question, the three assessors (raters) did sometimes select different items (categories) for the students' solutions (subjects).

	Assessor 1						Assessor 2						Assessor 3					
	S1	S2	S3	S4	S5	S6	S1	S2	S3	S4	S5	S6	S1	S2	S3	S4	S5	S6
(1)	X	X	X	X	X	X	X		X	X	X	X	X		X	X	X	X
(2)			X	X	X	X			X	X	X	X				X	X	X
(3)	X			X	X		X			X	X		X			X	X	X
(4)	X			X	X					X	X		X			X	X	X
(5)					X						X						X	
Score	3	1	1	3	2.5	1	2.5	0	1	3	2.5	1	3	0	1	3	2.5	3

Table 3.2 – Assessments by 3 assessors of 6 student's answers on the example question

Specification of the weight vector w

We start by specifying the weights of the vector w . The associated scores for each item will evidently play a crucial role in defining these. However, note that in Figure 3.2 the second (blue) item does not influence the final grade on the question. If our weights would only represent the associated scores, then $w_2 = 0$; meaning that item 2 would not play any role in the calculation of our kappa-statistic, while the presence/absence of the item changes the feedback a student receives. Hence, instead of using the (absolute value) of the associated score to define the weights, we add the maximum absolute value of the associated scores over all items. This means that the weights will be defined based on $|\text{score}_c| + \max_{k=1}^C \{|\text{score}_k|\}$. To get weights between 0 and 1, we divide this sum by the doubled maximum associated score over all items:

$$w_c = \frac{|\text{score}_c| + \max_{k=1}^C \{|\text{score}_k|\}}{2 \cdot \max_{k=1}^C \{|\text{score}_k|\}}. \quad (3.18)$$

These weights have a nice interpretation: the minimum weight is always 0.5, accounting for the (non-)selection of the item, everything between 0.5 and 1 depends on the

(absolute value of) the associated score of the item. As such, items that do not influence the final score, will have weight of 0.5, while items with the maximum (absolute value of) associated score will have weight 1. These weights do not sum to 1, considering their interpretation is more intuitive this way. Based on formula (3.18), the calculated weights for the example are given in Table 3.3.

Item		(1)	(2)	(3)	(4)	(5)
$ \text{score}_c $	(associated score)	1	0	1.5	0.5	0.5
$\max_{k=1}^C \{ \text{score}_k \}$	(selection)	1.5	1.5	1.5	1.5	1.5
Sum		2.5	1.5	3	2	2
Weight w_c		0.833	0.5	1	0.667	0.667

Table 3.3 – Specification of the weight vector w

Determining the matrix of possible classifications S and scale factors ϕ_c based on the hierarchical structure of the categories

We see that the first three items are all main categories: there are no conditions for (not) selecting them, so $s_{i1} = s_{i2} = s_{i3} = J = 3$ for every student i . For a possible classification into item 4, item 1 and item 3 must be selected first; for example, student 6 has only the third assessor selecting these, so $s_{6,4} = 1$. Item 5 can only be selected if item 4 was selected so $s_{i5} = x_{i4}$; for example, student 1 has 2 classifications for item 4 (assessor 1 & assessor 3), so $s_{1,5} = 2$. Matrix S can be found in Table 3.4.

The scale factors ϕ_c can be found by applying formula (3.16): for each category c , loop over all subject i and take the sum of the s_{ic} 's (sum up the columns of Table 3.4), and divide this sum by $IJ = 6 \cdot 3 = 18$.

S	s_{i1}	s_{i2}	s_{i3}	s_{i4}	s_{i5}
s_{1c}	3	3	3	3	2
s_{2c}	3	3	3	0	0
s_{3c}	3	3	3	0	0
s_{4c}	3	3	3	3	3
s_{5c}	3	3	3	3	3
s_{6c}	3	3	3	1	1
Sum	18	18	18	10	9
Scale factors	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5
	1	1	1	0.556	0.5

Table 3.4 – Determining the matrix of possible classifications S and scale factors ϕ_c

Calculating P_{O_c} and P_{e_c}

We give the full calculation of P_{O_1} and P_{e_1} in this paragraph. The other P_{O_c} 's and P_{e_c} 's can be calculated in a similar way. The required s_{i1} values were already calculated in

the previous step, we still need to count how many times item 1 was selected for each student i to get the x_{i1} values; the results can be found in [Table 3.5](#).

Student	S1	S2	S3	S4	S5	S6
x_{i1}	3	1	3	3	3	3
s_{i1}	3	3	3	3	3	3

Table 3.5 – Determining the x_{i1} 's and s_{i1} 's

Next, we calculate P_{O_1} based on formula (3.14):

$$P_{O_1} = \frac{[3 \cdot 2 + 0 \cdot -1] + [1 \cdot 0 + 2 \cdot 1] + [3 \cdot 2 + 0 \cdot -1] + [3 \cdot 2 + 0 \cdot -1] + [3 \cdot 2 + 0 \cdot -1] + [3 \cdot 2 + 0 \cdot -1]}{3 \cdot 2 + 3 \cdot 2 + 3 \cdot 2 + 3 \cdot 2 + 3 \cdot 2 + 3 \cdot 2} = 0.889,$$

For the computation of P_{e_1} , we use formula (3.15):

$$P_{e_1} = 2 \cdot \left(\frac{3 + 1 + 3 + 3 + 3 + 3}{3 + 3 + 3 + 3 + 3 + 3} \right)^2 - 2 \cdot \left(\frac{3 + 1 + 3 + 3 + 3 + 3}{3 + 3 + 3 + 3 + 3 + 3} \right) + 1 = 0.802,$$

Although not necessary for the calculation of our proposed κ statistic, it is possible to calculate the partial κ_c to have an indication of the reliability of each item. For item 1, this becomes (see formula (3.5)):

$$\kappa_1 = \frac{P_{O_1} - P_{e_1}}{1 - P_{e_1}} = \frac{0.889 - 0.802}{1 - 0.802} = 0.438.$$

Although item 1 was selected for most students (only assessor 2 and 3 did not select it for student 2), we get a relatively low κ_1 -value. How can this be explained? Item 1 was chosen for almost all students by almost all assessors, leading to a high agreement by chance P_{e_1} ($= 0.802$). This means that without even looking at a student's solution, there is a high probability that an assessor selects item 1. The fact that student 2 has two non-classifications for item 1 while assessor 1 did select item 1 for this student leads, therefore, leads to a pretty severe penalisation in the partial kappa κ_1 . This is a concrete example of the 'prevalence paradox' described in [section 3.1.3](#).

The other P_{O_c} 's and P_{e_c} 's can be calculated analogously. The result can be found in [Table 3.6](#).

Items	(1)	(2)	(3)	(4)	(5)
P_{O_c}	0.889	0.889	0.889	0.778	1.00
P_{e_c}	0.802	0.525	0.506	0.820	0.556
$P_{O_c} - P_{e_c}$	0.086	0.364	0.383	-0.042	0.444
$1 - P_{e_c}$	0.198	0.475	0.494	0.180	0.444
κ_c	0.438	0.766	0.775	-0.235	1.00

Table 3.6 – P_{O_c} , P_{e_c} , $P_{O_c} - P_{e_c}$, $1 - P_{e_c}$ and partial kappa κ_c for every item (=category)

Calculation of the kappa-statistic

With the specification of weight vector \mathbf{w} , and the computation of the scale factors ϕ_c , the ‘beyond-chance’ $P_{O_c} - P_{e_c}$ and the ‘beyond-chance in case of perfectly agreeing raters’ $1 - P_{e_c}$, we are ready to calculate the kappa-statistic for the hierarchical case (see formula (3.17)):

$$\begin{aligned} \kappa &= \frac{0.833 \cdot 1 \cdot 0.086 + 0.5 \cdot 1 \cdot 0.364 + 1 \cdot 1 \cdot 0.383 + 0.667 \cdot 0.556 \cdot (-0.042) + 0.667 \cdot 0.5 \cdot 0.444}{0.833 \cdot 1 \cdot 0.198 + 0.5 \cdot 1 \cdot 0.475 + 1 \cdot 1 \cdot 0.494 + 0.667 \cdot 0.556 \cdot 0.180 + 0.667 \cdot 0.5 \cdot 0.444} \\ &= 0.692. \end{aligned}$$

We get a relatively high κ -value, that would be labelled by the benchmark scale of Landis and Koch (1977) as ‘Substantial’ agreement.

3.3.1.3 Comparison with other methods

We also calculated this example through the other methods described in section 3.1.4. Averaging/pooling Cohen’s kappas is no possibility as we have more than two raters. The proportional overlap method is possible and returns $\kappa = 0.602$. However, the method is based on some questionable premises in this context: (1) it assumes all items are equally weighted (so there is no correction for the associated scores), (2) it assumes all categories are always available to all raters (so the hierarchy of the items is ignored). Besides, the method fails to measure potential observed agreement for student 2 as $A_{22} = A_{23} = \emptyset$, no proportional overlaps can be calculated. Problems (1) and (2) also occur with the chance-corrected intraclass correlations that return a κ -value of 0.379. The problem of failing to measure potential observed agreement for student 2 emerges in another guise: while the proportional overlap method leaves student 2 out of the calculation of P_o , the chance-corrected intraclass correlations do include student 2 with an intraclass correlation coefficient of almost zero, pulling down the P_o value in an unacceptable way. While our proposed κ statistic entails the philosophical premise that two raters not selecting category c is equally valued in terms of agreement as two raters who do select category c ; these examples show that the opposite - completely exclude agreement in non-selections - also can lead to unsatisfactory results. Finally, the calculation of chance-corrected rank correlations are not relevant in this context as raters do not make ordered classifications in checkbox grading.

3.3.2 Diagnosing psychiatric cases

We now revisit an example from Mezzich et al. (1981). It consists of a diagnostic exercise in which 30 child psychiatrist made independent diagnoses of 27 child psychiatric cases. Each psychiatrist rated 3 cases, and each case turned out to be rated by 3 or 4 psychiatrists upon completion of the study. Table 3.7 shows the 90 multiple diagnostic formulations. Each diagnostic formulation presented was composed of up to three from the twenty broad diagnostic categories taken from Axis I (clinical psychiatric syndromes) of the American Psychiatric Association’s Diagnostic and Statistical Manual of Mental Disorders (DSM-III). We are well aware that DSM-III is outdated (American Psychiatric Association, 2022), but the example remains excellent as it can be contrasted with the other measures in the literature.

Table 3.7 – Multiple diagnostic formulations from 27 child psychiatric cases using DSM-III Axis I Broad Categories*

Cases	Raters			
	1	2	3	4
1	9, 11	11, 9, 14	16, 9	11, 9
2	16	16, 14	12	14, 5
3	17	12	7, 8	13
4	16, 13	13, 16, 14	16	
5	7	7, 12, 13	13	
6	10	10	10	
7	7, 16	13	16	
8	1, 14	13	16, 13	
9	5	20	13, 14	
10	12, 13, 14	12, 14, 13	12, 11 14	
11	13	18	16	
12	5, 18	1, 5, 18	1	
13	14, 13	14, 7	14, 16	
14	11, 16	14, 11, 16	11, 13	
15	10	3, 18	10, 11	
16	14, 5	5, 16	14	
17	12	12, 11	12	
18	20	16	16	
19	13	14	14	
20	9, 14, 10	9, 11, 14	10, 9	
21	12, 11	11, 14	11	
22	17	12	12	12, 17, 15
23	16, 13	12	14	13
24	12	12	16	12
25	13	20	13	13
26	13	13, 16	13	16
27	10, 9	9, 10	9	9, 10

* 1. Organic mental disorders, 2. Substance use disorders, 3. Schizophrenic and paranoid disorders, 4. Schizoaffective disorders, 5. Affective disorder, 6. Psychoses not elsewhere classified, 7. Anxiety factitious, somatoform and dissociative disorders, 8. Pyschosexual disorder, 9. Mental retardation, 10. Pervasive developmental disorder, 11. Attention deficit disorders, 12. Conduct disorders, 13. Anxiety disorders of childhood or adolescence, 14. Other disorders of childhood or adolescence, speech and stereotyped movement disorders, disorders characteristic of late adolescence, 15. Eating disorders, 16. Reactive disorders not elsewhere classified, 17. Disorders of impulse control not elsewhere classified, 18. Sleep and other disorders, 19. Conditions not attributable to a mental disorder, 20. No diagnosis on Axis I.

We start with the calculation of our proposed κ statistic. The example consists of 27 child psychiatric cases (i.e., the number of subjects I equals 27), to be classified into 20 broad diagnostic categories (i.e., the number of categories C equals 20) with a varying number of raters, expressed in vector \mathbf{j} with $j_i = 3$ or $j_i = 4$, depending on the case, see Table 3.8.

Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
j_i	4	4	4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4

Table 3.8 – Number of psychiatrists (= raters) for each case i (= subject)

We assume all diagnostic categories are equally important and thus use unweighted categories ($w_c = 1, \forall c$). Moreover, the diagnostic categories on Axis I have no hierarchy. Hence, we can use the formulas described in section 3.2.1. First, we calculate matrix X by counting how many times a diagnostic category c appeared for a subject i (e.g., $x_{1,1} = 0, x_{12,1} = 2, x_{6,10} = 3, \dots$). Next, we combine the x_{ic} 's and the j_i 's to determine the P_{O_c} 's (3.12) and the P_{e_c} 's (3.13). As an example, we calculate P_{O_1} and P_{e_1} :

$$P_{O_1} = \frac{9[2 \cdot 0^2 - 2 \cdot 4 \cdot 0 + 4^2 - 4] + 16[2 \cdot 0^2 - 2 \cdot 3 \cdot 0 + 3^2 - 3] + 1[2 \cdot 1^2 - 2 \cdot 3 \cdot 1 + 3^2 - 3] + 1[2 \cdot 2^2 - 2 \cdot 3 \cdot 2 + 3^2 - 3]}{9[4(4-1)] + 16[3(3-1)] + 1[3(3-1)] + 1[3(3-1)]}$$

$$= 0.963$$

$$P_{e_1} = 2 \left(\frac{9 \cdot 0 + 16 \cdot 0 + 1 \cdot 1 + 1 \cdot 2}{9 \cdot 4 + 16 \cdot 3 + 1 \cdot 3 + 1 \cdot 3} \right)^2 + 2 \left(\frac{9 \cdot 0 + 16 \cdot 0 + 1 \cdot 1 + 1 \cdot 2}{9 \cdot 4 + 16 \cdot 3 + 1 \cdot 3 + 1 \cdot 3} \right) + 1$$

$$= 0.936.$$

The other calculations can be found in Table 3.9.

Diagnose	1	2	3	4	5	6	7	8	9	10
P_{O_c}	0.963	1.000	0.981	1.000	0.917	1.000	0.917	0.972	1.000	0.935
P_{e_c}	0.936	1.000	0.978	1.000	0.876	1.000	0.895	0.978	0.785	0.802
$P_{O_c} - P_{e_c}$	0.027	0.000	0.003	0.000	0.041	0.000	0.022	-0.006	0.215	0.133
$1 - P_{e_c}$	0.064	0.000	0.022	0.000	0.124	0.000	0.105	0.022	0.215	0.198
κ_c	0.425	NaN	0.157	NaN	0.330	NaN	0.206	-0.264	1.000	0.672
Diagnose	11	12	13	14	15	16	17	18	19	20
P_{O_c}	0.898	0.824	0.694	0.759	0.972	0.713	0.935	0.944	1.000	0.935
P_{e_c}	0.753	0.694	0.620	0.642	0.978	0.654	0.936	0.915	1.000	0.936
$P_{O_c} - P_{e_c}$	0.145	0.130	0.075	0.117	-0.006	0.059	0.000	0.029	0.000	0.000
$1 - P_{e_c}$	0.247	0.306	0.380	0.358	0.022	0.346	0.064	0.085	0.000	0.064
κ_c	0.588	0.426	0.197	0.327	-0.264	0.170	-0.006	0.346	NaN	-0.006

Table 3.9 – $P_{O_c}, P_{e_c}, P_{O_c} - P_{e_c}, 1 - P_{e_c}$ and partial kappa κ_c for every diagnostic category

Note that $\kappa_2, \kappa_4, \kappa_6$ and κ_{19} equal NaN, due to a division by zero. Such division by zero will always happen if no rater chooses a category. As $P_{O_c} = P_{e_c} = 1$ in those categories and thus $P_{O_c} - P_{e_c} = 1 - P_{e_c} = 0$, these unchosen categories do not play any role in the calculation of the proposed κ statistic. Hence, the κ statistic is independent of unused alternative categories, meaning it can not be inflated by adding unchosen categories;

we get from formula (3.8):

$$\kappa = \frac{\sum_c (P_{o_c} - P_{e_c})}{\sum_c (1 - P_{e_c})} = 0.375.$$

We get a relatively low kappa-value, which should not come unexpected: [Table 3.7](#) shows that the various psychiatrists diverge rather vehemently in their diagnoses. The proposed κ statistic yields a higher value than the proportional overlap method ($\kappa = 0.27$), but is almost equal to the chance-corrected intraclass correlation method ($\kappa = 0.35$) and the rank correlation method ($\kappa = 0.34$).

3.4 FURTHER RESEARCH

The story of the proposed κ statistic is not finished by this chapter. First, the publication of an R package is envisioned containing ready-to-use functions to calculate all described measures. Such a package would allow researchers without an overly statistical background to use the measure in their research and can greatly facilitate the adoption of the proposed measure.

In addition, more can be told about the proposed measure. Based on (De Vries et al., 2008), we envision publishing the simulation study to show that our proposed kappa statistic exhibits smaller root-mean-square errors than taking a weighted average of Fleiss' kappas. Moreover, the large-sample variance of the proposed κ statistic still needs to be determined. An expression for the variance would enable statistical inference using the measure without bootstrapping. It especially paves the way for performing robust power analysis: researchers wishing to set up an experiment in which raters classify subjects into one-or-more categories would be able to calculate in advance the number of raters and subjects required to reach a certain confidence level. Finding the large-sample variance of our proposed κ statistic is by no means an easy quest: it took the scientific community 50 years to develop a general expression for the Fleiss' kappa! Indeed, it was Gwet (2021) who finally came up with a correct formula for the variance of the Fleiss' kappa. The variance described in Fleiss (1971) is simply wrong; the standard error of Fleiss et al. (1979) is valid only under the assumption of no agreement among raters; as such, it can only be used to test the hypothesis of zero agreement among the raters. Unfortunately, as many statistical software programs provide the standard error of Fleiss et al. (1979) along with the calculation of Fleiss' kappa, it is immensely misused for all kinds of statistical inference. Let us avoid making the same mistakes when searching a large-sample variance of our proposed measure that presumably entails a generalisation of the formula found by Gwet (2021).

Finally, now that we have established the idea of the proposed κ statistic, the same idea may be suitable to create other long-needed measures. For example, the literature on rubrics (Dawson, 2017) lacks a unified way to compare the inter-rater reliability of two rubrics assessing the same phenomenon (e.g. book reviews of students, PhD proposals). Should such a measure exist, it would be possible to compare the impact of including/excluding specific criteria. Such a measure can possibly be constructed by the calculation of the P_o and P_e of the Fleiss kappa (or the Krippendorff's alpha,

see Gwet, 2012) for groups of criteria assessing the same aspect and weighting them according to the maximum score of the aspect.

3.5 CONCLUSION

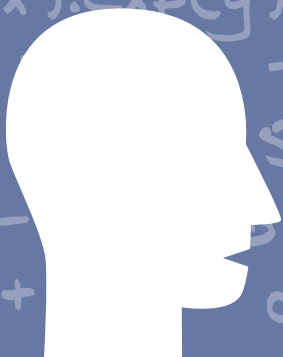
This chapter has presented a generalisation of Fleiss' kappa, allowing raters to select multiple categories for each subject. Categories can be weighted according to their importance in the research context, and the measure can account for possible hierarchical dependencies between the categories. A crucial assumption of the proposed κ statistic is that two raters selecting a specific category for a given subject count equally in agreement as two raters not selecting the category. Other methods, like proportional overlap, chance-corrected intraclass correlations and chance-corrected rank correlations, do not make this assumption; instead, they ignore the agreement in the non-selection of categories. We have shown that this ignorance can give unexpected and unwanted results depending on the research context. By introducing this generalisation of Fleiss' kappa and comparing and contrasting it to the existing comparable methods, we hope to inspire further researchers in need of a chance-corrected inter-rater reliability measure that allows measuring the agreement among several raters classifying subjects into one-or-more (hierarchical) nominal categories.

CRedit authorship contribution statement

Filip Moons: Conceptualisation, Formal analysis, Investigation, Project administration, Writing – original draft. ● **Ellen Vandervieren:** Writing – review & editing, Supervision, Funding acquisition.

HIGHLIGHTS

- ✔ A study conducted with the Flemish Exam Commission exploring the use of 'checkbox grading' from the assessors' perspective. Assessors receive a list of checkboxes and must select those that apply to the student's solution. Checkbox dependencies can be set to ensure consistent grading. The system then calculates the grade and provides atomic feedback.
- ✔ Checkbox grading took twice as long as traditional grading but had equally high inter-rater reliability. Paradoxically, assessors felt that checkbox grading allowed them to complete their tasks more quickly.
- ✔ Checkbox grading was investigated in two conditions: the blind and visible condition. In the blind condition, assessors could not see any grades, while they could in the visible condition. The blind condition enhanced inter-rater reliability when grading schemes should be interpreted strictly.
- ✔ Assessors unanimously favoured visible over blind checkbox grading. Seeing the grades helps them judge the correctness of their assessment work, and they miss having that opportunity in blind grading.



☑ Chapter 4

CHECKBOX GRADING OF HANDWRITTEN MATHEMATICS EXAMS WITH MULTIPLE ASSESSORS: STUDY ON TIME, INTER-RATER RELIABILITY, USAGE & VIEWS

PUBLICATIONS

Moons, F., Vandervieren, E. & Colpaert, J. (Under review). Checkbox grading of handwritten mathematics exams with multiple assessors: study on time, inter-rater reliability, usage & views. *Journal for Research in Mathematics Education*.

Moons, F., & Vandervieren, E. (2023). Blind versus visible checkbox grading: does not seeing the grades when assessing mathematics enhance inter-rater reliability? *13th Congress of European Research in Mathematics Education (CERME13)*, 10-14 July 2023 in Budapest, Hungary.

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)*, 2-5 February 2022 in Bozen-Bolzano, Italy. <https://hal.science/hal-03753446>

ABSTRACT

Digital exams often fail to assess all required mathematical skills. Therefore, it is advised that large-scale exams still feature some handwritten open-answer tasks. However, assessing those handwritten tasks with multiple assessors is often challenging regarding inter-rater reliability and feedback. This chapter focuses on the scoring process of handwritten mathematics exam tasks and presents a new semi-automated approach called 'checkbox grading.' Exam designers predefine atomic feedback items with partial grades; next, assessors should just tick the items ('checkboxes') relevant to a student's answer. Dependencies between these items can be set to ensure that every assessor takes the same path down the grading scheme. Moreover, the approach allows 'blind checkbox grading' where the underlying grades are not shown to the assessors. The approach was studied during a large-scale advanced mathematics exam of the senior year of compulsory education (12th grade), organised by the Flemish Exam Commission. Results show that assessors perceived checkbox grading as very useful and had a high attitude towards using it. However, compared to traditional grading – just following a correction scheme and communicating the resulting grade – the time investment for assessors is higher, while both approaches are equally reliable. Nevertheless, blind checkbox grading improved inter-rater reliability for some exam tasks. Altogether, checkbox grading might lead to a smoother correction process for paper-and-pencil tasks in which feedback can be communicated to students, not solely grades.

4.1 INTRODUCTION

Regardless of all the practical advantages digital exams offer, Hoogland and Tout (2018) warn that digital mathematics tasks often focus on lower-order goals (e.g., procedural skills). They argue that handwritten tasks 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 task is asked digitally or paper-based. Bokhove and Drijvers (2010) point out that handwritten tasks allow students to express themselves more freely. In a more recent study, Backes and Cowan (2019) conducted the same test with two groups: one took the test digitally, while the other used the traditional paper-and-pencil format. Notably, the digital group exhibited downward trends that could not be accounted for by pre-existing tendencies. For all these reasons, 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).

Major issues with assessing handwritten tasks in large-scale exams are finding efficient ways to provide consistent feedback and to assess reliably when multiple assessors are involved (Baird et al., 2004; Meadows & Billington, 2005). *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).

Therefore, in some countries like the United Kingdom and the Netherlands, national exams are assessed by two assessors to guarantee judgment accuracy (Brooks, 2004; Kuhlemeier et al., 2012). Interestingly, studies on grading reliability in mathematics education can be traced back over a century, as demonstrated by Starch and Elliott's (1913) investigation, where geometry exams were sent to mathematics teachers in 1912. Notably, the grades varied considerably due to the absence of uniform grading criteria. Nowadays, most exam designers try to ensure inter rater-reliability by pre-developing a solution key with grading instructions for assessors (Ahmed & Pollitt, 2011).

However, pre-developed grading instructions are not perfect. One source of assessor variability emerges when the assessors' *holistic* grade differs from the *calculated* grade, which is repeatedly described in the literature on rubrics (Huot, 1990; Moskal & Leydens, 2000; Stellmack et al., 2009). The holistic grade is what assessors intuitively want to give when scoring a student's product (e.g., a math exam task). In contrast, the calculated grade is obtained by deliberately following the scoring guidelines from the rubric criteria. When the calculated grade does not align with the holistic evaluation of the work, assessors often start changing the selection of criteria, which compromises the instrument's reliability (Dawson, 2017). This cognitive conflict between holistic and calculated grades links to the well-known dual process theory from cognitive psychology (Evans, 2008).

This chapter introduces 'checkbox grading': an assessment method for handwritten mathematics exam tasks. It investigates how assessors used the method to evaluate 60 students' 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 by reducing the assessor variability. Moreover, the grading method results in feedback for the students, giving a detailed insight into how grades were obtained. However, our focus in this chapter lies on the assessors' perspectives rather than those of the students.

In the following sections, we discuss this method, the idea of 'blind' grading, introduce the research framework and state the research questions.

4.1.1 Checkbox grading

Checkbox grading is a semi-automated method to assess handwritten mathematics tasks: students solve tasks the classical way by writing on a sheet of paper. Next, these sheets are scanned, and assessors use the checkbox grading system to correct the solutions on a computer. The exam designers provide a grading scheme for each task consisting of different feedback items written in an *atomic way* (see below), anticipating most mistakes. These feedback items can be linked to partial points for grading. When correcting a student's solution, the assessors must just select the feedback items ('checkboxes') that apply (see Figure 4.1). Dependencies between these items can be set to ensure that every assessor takes the same path down the grading scheme. When all assessors finish their job, the system produces individual student reports, including the grades and feedback (like in Figure 4.1).

The name 'checkbox grading' was inspired by the bestseller 'The Checklist Manifesto' by Gawande (2009), in which the author argues that using simple checklists in daily and professional life can make even very complex processes efficient, consistent and safe:

(/2.5) Calculate $\frac{1+3i}{-2-5i}$ and write the answer in a+bi form.
Show all your intermediate steps, don't use your calculator.

Student's answer

$$\frac{1+3i}{-2-5i} = \frac{-1-3i}{-2-5i} \cdot \frac{(-2+5i)}{(-2+5i)}$$

$$= \frac{(-1-3i)(-2+5i)}{4-25i^2} = \frac{-15i^2-5i+6i+2}{29}$$

$$= \frac{-17+6i-5i}{29}$$

Solution key

$$\frac{1+3i}{-2-5i} = \frac{1-3i}{-2-5i}$$

$$= \frac{(1-3i) \cdot (-2+5i)}{(-2-5i)(-2+5i)}$$

$$= \frac{-2+5i+6i-15i^2}{4+25} = \frac{13+11i}{29}$$

$$= \frac{13}{29} + \frac{11}{29}i$$

Correction by assessor

🔍 First check-up

- No intermediate steps provided **max: 0.0**
- Solved using the *polar form of complex numbers* which is impossible without calculator **max: 0.0**

! Checking the calculation

- Correct complex conjugate $1 - 3i$ in the numerator. **+0.5**
- If the complex conjugate in the numerator is miscalculated or not applied, the student's answer will deviate from the solution key. Therefore, it is necessary to check the student's calculation individually for the indicated items.
 - Check individually: Correctly multiplied by the conjugate binomial in the denominator **+0.5**
 - Denominator may also be calculated immediately (= 29)
 - $-(2 - 5i)$ is also fine (denominator in this case = -29)
 - Also fine if more steps were used (e.g., first $-(2 + 5i)$, next $-(21 + 20i)$)
 - Check individually: Correct calculation of the numerator with intermediate step **+0.5**
 - Correct denominator (=29 or =-29) **+0.5**
 - Correct final answer in $a + bi$ form **+0.5 if calculation is fully correct**

Grade: 1/2.5

Figure 4.1 – Checkbox grading scheme of exam task 1

“under conditions of complexity, not only are checklists a help, they are required for success.”

4.1.1.1 Adaptive grading

To obtain grades using checkbox grading, exam designers can associate items with partial points to be added (green items in Figure 4.1) or subtracted. It is also possible to associate items with a threshold (e.g., ‘if this feedback item is ticked, no points’, red items in Figure 4.1).

The point-by-point list of atomic feedback items ultimately forms a series of implicit yes/no questions to determine the students’ grade. The dependencies that can be set

between items consist of showing, disabling or changing items whenever a previous item is ticked, implying that assessors must follow the point-by-point list from top to bottom. This adaptive grading approach resembles a flowchart that automatically determines the grade, but – by ticking the items that are relevant to a student's answer – might at the same lead to several other envisioned benefits: (1) a deep insight into how the grade was obtained for both the student (feedback) as well as the exam designers, and (2) a straightforward way to do correction work with multiple assessors as personal interpretations are avoided as much as possible (inter-rater reliability).

In [Figure 4.1](#), an example of this approach is given. The student's answer survives the 'First check-up' items; checking one of them would otherwise disable all of the following items. As the item 'Correct complex conjugate $1 - 3i$ ' is unticked, the computer knows that a mistake happened; however, assessors should continue their assessment of the answer by taking into account that the students' steps will now deviate from the solution key for some items; these items are indicated by 'Check individually.' All the orange content would have disappeared if the item 'Correct complex conjugate $1-3i$ ' had been ticked. The item 'Correct final answer in $a + bi$ form' only gets enabled when all previous green items are ticked. The two ticked items each add 0.5 points to the grade, leading to a total of 1 out of 2.5.

4.1.1.2 Atomic feedback

When developing a checkbox grading scheme, exam designers must anticipate the mistakes that students can make. In doing so, the exam designers should balance the grading scheme's rigidity and how explicitly they want to address certain mistakes while ensuring the checkbox grading items are as reusable as possible. To tackle these challenges, the exam designers are encouraged to write the feedback items in an *atomic* way. Atomic feedback is a set of form requirements from which it is shown that it makes feedback significantly more reusable (Moons et al., 2022; see [Chapter 1](#)). To write atomic feedback¹, one has to (1) identify the possible independent errors occurring and (2) write separate feedback items for each error, independent of each other. Since the atomic items are shared across multiple assessors in this second study, an additional criterion for atomicity is added: (3) a knowledgeable assessor must be able to determine unambiguously whether an item applies to a student's answer or not. As such, each item implicitly represents a yes/no question. These three rules guide the development of a checkbox grading scheme, and the level of detail of the checkbox grading scheme is tightly related to how atomic the feedback items are formulated. The 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 [Figure 4.1](#)); moreover, related atomic feedback items and intermediate steps in a solution key can share the same colour to make their connection visually clear (see [Figure 4.1](#)).

Suppose a particular solution approach by a student is not covered in the available feedback items. In that case, an assessor can add a new feedback item, leading to a dynamic grading scheme that expands as more and more exams get graded. Practically,

¹We only present the definition of atomic feedback in this study and not the more detailed guidelines of [Appendix D](#), as these guidelines were used to differentiate individual teachers' feedback items into atomic/non-atomic.

The screenshot shows a list of grading criteria with checkboxes and icons for editing or deleting. The criteria are:

- Check individually: Correct calculation of the numerator with intermediate step (-
- Correct denominator (=29 of =-29) (+0.5) ↓ → ✕
- Correct final answer in $a + bi$ form (+0.5 if calculation is fully correct) ↓ → ✕
- I'm adding some item |

Figure 4.2 – An example of a dynamic grading scheme where an assessor can add items for unforeseen solution approaches

the assessors should tap on the arrow where they would like to add a feedback item, they write it, and it is immediately available to the other assessors (see [Figure 4.2](#)). The exam designers can make the added item an official part of the grading scheme by assigning partial points to it. The grades of the solutions where the item was selected are automatically recalculated in that case. However, revising the already assessed exams is unnecessary since, theoretically, there was no need for the added item at the time.

4.1.1.3 Blind grading

Imagine that all references to partial scores and the final grade disappear in [Figure 4.1](#). This leads to the experimental idea of ‘blind grading’ where the assessor chooses the appropriate feedback items without seeing the associated scores. The system still calculates the grades, but these are invisible to the assessors. The envisioned advantage of this grading approach is that assessors only need to focus on the content of a student’s answer; any emotional barrier to selecting a feedback item disappears, possibly leading to higher grading reliability, as it avoids the cognitive conflict between holistic and calculated grades.

The opposite mode of blind grading will be called ‘visible grading’ in the rest of the chapter; this is the standard mode where assessors can see the associated points for every feedback item and the calculated total grade (see [Figure 4.1](#)).

Note that blind grading should not be confused with anonymous grading (Hanna & Linden, 2012); in anonymous grading, assessors do not see the students’ names to avoid certain biases (e.g., gender, ethnicity).

4.2 RESEARCH FRAMEWORK & QUESTIONS

The checkbox grading approach was developed in cooperation with the Flemish Exam Commission. In this study, *traditional grading* serves as a benchmark for the checkbox grading approach. Traditional grading is the usual procedure of the Flemish Exam Commission to assess handwritten mathematics tasks: assessors receive a PDF file with grading guidelines for every exam task (see [Figure 4.3](#) to see the traditional grading scheme of the task shown in [Figure 4.1](#)), have access to the scanned students’ exams and only communicate a grade for each task based on the guidelines.

In this study, we want to investigate the assessors’ time investment and inter-rater reliability to compare blind with visible checkbox grading on the one hand and checkbox

No points when no intermediate steps/solution method provided. Can't be solved by using the polar form of complex number because must be solved without calculator!

$$\frac{1+3i}{-2-5i} = \frac{1-3i}{-2-5i}$$

- 0,5 point (or 0) for correct complex conjugate in the numerator
- NOTE: if not applied or wrong: follow through with mistake, max 1.5/2.5 because: 0/0,5 for correct complex conjugate in the numerator 0/0,5 for correct final answer (last step)

$$= \frac{(1-3i) \cdot (-2+5i)}{(-2-5i)(-2+5i)} (*)$$

- 0,5 point (or 0) for multiplication with the conjugate binomial in the denominator
- NOTE: denominator may be calculated immediately (=29)
- NOTE: (.2-5i) is also fine (denominator in this case = -29)
- NOTE: also fine if more steps were used (e.g. first, (.2+5i), next (.21+20i))
- NOTE: if binomial conjugate is wrong or missing, no points for the rest of the student's solution

$$= \frac{-2+5i+6i-15i^2}{4+25} = \frac{13+11i}{29}$$

- 0,5 point (or 0) for correct calculation of the numerator with intermediate step
- 0,5 point (or 0) for correct denominator (=29 or = -29)
- 0,5 point (or 0) for correct final answer in a + bi form if obtained from (*) with at least 1 intermediate step

$$= \frac{13}{29} + \frac{11}{29}i$$

Figure 4.3 – Traditional grading scheme of exam task 1

grading with traditional grading on the other hand. To get a grip on the assessors' experiences, we also investigate their views and usage regarding checkbox grading.

To frame our research, we use the DiaCoM framework (Loibl et al., 2020). The DiaCoM framework is a cognitive model describing diagnostic judgments when assessing students' performance. The framework distinguishes between the *situation characteristics* (e.g., the framing of the assessment: grade level, importance,... and the cues available to assessors like tasks, guidelines) and *person characteristics* (e.g., assessors' states and traits). The backbone of the framework is the diagnostic process and is split into two components: the internal information processing consisting of *diagnostic thinking* and the *diagnostic behaviour*, which constitutes the external verbalisation of the assessor. *Diagnostic thinking* describes the cognitive thinking process during the genesis of a diagnostic judgment and is split into perceiving, interpreting, and decision-making. *Diagnostic behaviour* is split into process and product indicators.

The DiaCoM framework for our research is displayed in Figure 4.4. The study did not influence the *person characteristics* in any way. Concerning the *situation characteristics*, the study investigates a high-stakes exam organized by the Flemish Exam Commission at the end of grade 12 and graded by assessors who do not know the students. The foundation of our study is changing the cues of assessors to perform a diagnostic judgement by switching between visible/blind checkbox grading & traditional grading. By doing so, we influence their *diagnostic thinking* which consists of three activities: assessors perceive the empty checkbox/traditional grading schemes, interpret the student's solution, and decide which checkboxes apply (blind/visible checkbox grading) or the resulting grade from the traditional grading schemes (traditional grading). We

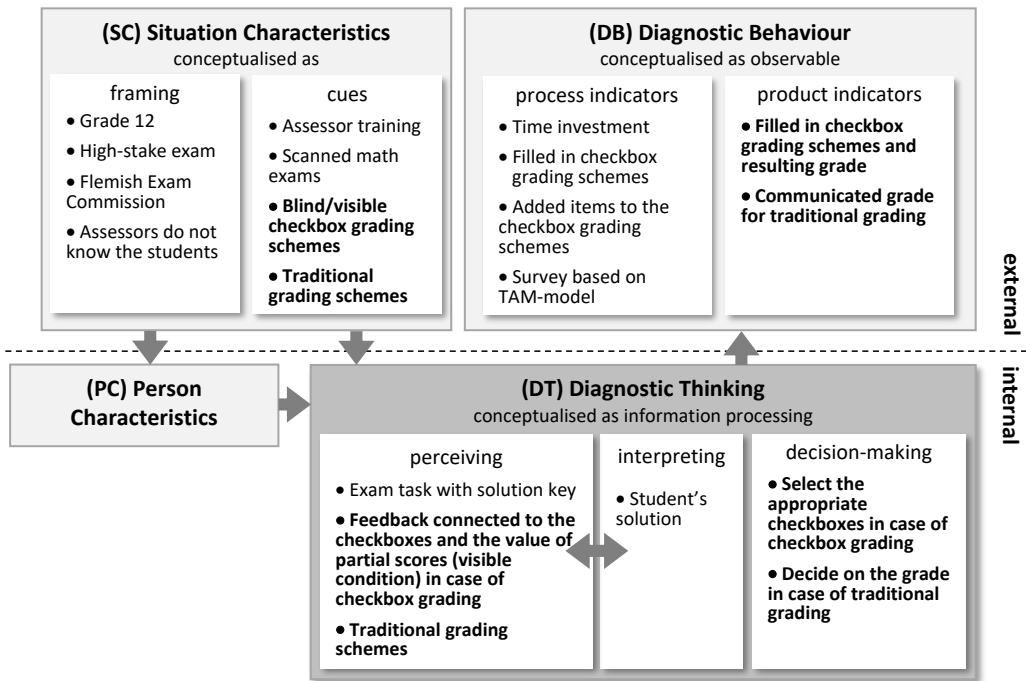


Figure 4.4 – The DiaCoM framework (Loibl et al., 2020) applied to our study

opted to strictly separate perceiving the checkbox/traditional grading schemes from interpreting the students' solution, as the study aims to conceive a grading method that is as objective as possible when multiple assessors are involved. As such, personal interpretations of the grading schemes by individual assessors should be avoided to the maximum extent possible. Obviously, this separation is an idealisation of the real cognitive process in which perceiving the grading schemes and interpreting the student's solution will inevitably affect each other, hence the added double arrow. The possible cognitive conflict between holistic and calculated grades also resides in this interplay when the grades are shown (in visible checkbox grading and traditional grading). The decisions made during the *diagnostic thinking* process also serve as product indicators of assessors' *diagnostic behavior*. As process indicators, we measured assessors' time spent in all conditions, their survey answers, and the filled-in checkbox grading schemes and possibly added items (see Figure 4.2). Note that checkbox grading gives an insight into both the process and the product, while traditional grading only provides product indicators.

Following this framework, this study formulates three research questions to investigate the assessors' time investment, inter-rater reliability, usage, and views:

- [RQ 4.1] Is there a difference in time investment between blind and visible checkbox grading? And between checkbox grading and traditional grading?
- [RQ 4.2] Does inter-rater reliability differ between blind and visible checkbox grading? And between checkbox grading and traditional grading?
- [RQ 4.3] How did assessors use and perceive checkbox grading?

4.3 METHODS & MATERIALS

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.

The study was conducted with the Secondary Education Exam Commission of Flanders (the Dutch-speaking part of Belgium). Flanders is a region without any central exams (Bolondi et al., 2019): every secondary school decides autonomously on the assessment of students. Consequently, the Exam Commission does not organise national exams for all Flemish students. However, it organises large-scale 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. Students participating in these exams prepare by self-study or using a private tutor/school. The commission provides clear guidelines for students on the content of the exams, carries out all the exams, and awards diplomas; but does not provide any teaching activities or materials to students.

A timeline of the study can be found in [Figure 4.5](#), the experiment started the 1st of October, 2021. In the subsequent sections, we will discuss each step along this timeline.

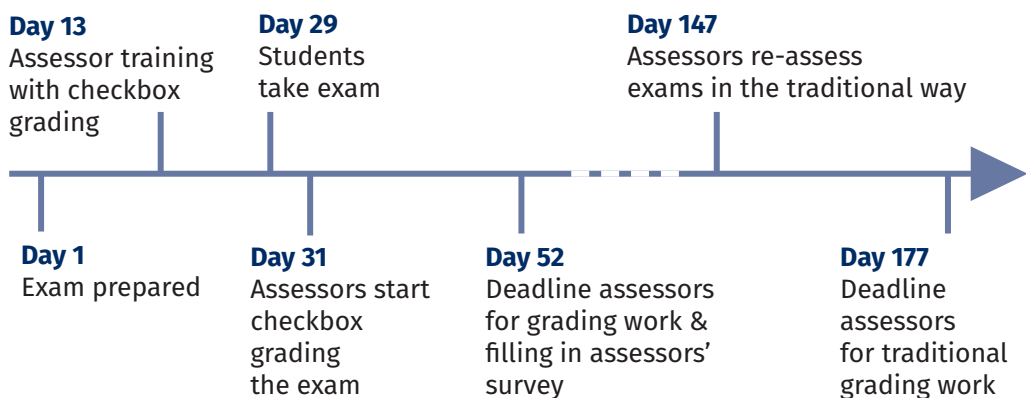


Figure 4.5 – Timeline of the study

4.3.1 Exam & checkbox grading tool development

The mathematics exam for this study was developed by the three mathematics exam designers of the Flemish Exam Commission. The exam was one of the two math exams for the advanced mathematics track of the senior years of Flemish secondary education (11th/12th grade). It featured complex numbers, matrices, solid geometry, discrete mathematics, statistics, and probability. Interestingly, the exam was already a mixture of fully automated and paper-and-pencil tasks: 46% of the exam grades were obtained with digital exam tasks (e.g., multiple choice and closed answer tasks). Our study will only focus on the 54% that consists of 10 paper-based exam tasks with an open-answer format: whenever we refer to the exam (results), we refer to this paper-based part. An overview of the exam tasks of the paper-based exam can be found in [Table 4.1](#). The

#	Topic	Learning goal	Max. score	Avg. score M ± SD
T1	Complex numbers	Calculations with complex numbers in $a + bi$ <-form	2.5	1.75 ± 0.88
T2	Complex numbers	Calculations with complex numbers in polar form	2.5	0.67 ± 0.61
T3	Matrices	Modelling with matrices	3.5	1.95 ± 0.96
T4	Matrices	Coefficient matrices of linear equations	3.5	1.18 ± 0.96
T5	Solid geometry	Parameter equations of a plane	1.5	0.18 ± 0.42
T6	Solid geometry	Cartesian equation of a line	1	0.04 ± 0.20
T7	Solid geometry	Drawing a line segment in the x,y,z axis system	2.5	1.16 ± 0.80
T8	Solid geometry	Determining the distance between a point and a line	4.5	0.57 ± 1.35
T9	Solid geometry	Parallel lines in solid geometry	2.5	0.76 ± 0.94
T10	Statistics & Probability	Modelling a probability experiment	4	0.39 ± 1.13
Total			28	8.65 ± 4.93

Table 4.1 – Content of the mathematics exam, including the scores' maximum, mean and standard deviation

tasks vary considerably in points that could be gained, based on the importance of the topic in the curriculum and the complexity of the task; 0.5 points was the smallest possible partial score. The exam development took place without any influence from the researchers and in the way they always develop exams, resulting for the paper-based exam in tasks with traditional grading schemes.

After the exam designers finished creating the exam, their traditional grading schemes were turned into checkbox grading in close cooperation with the researchers. For example, the traditional grading guidelines established by the exam designers of task 1 are shown in [Figure 4.3](#). By comparing [Figure 4.3](#) with the checkbox grading version in [Figure 4.1](#), one can get an idea of how this transformation took place. We ensured that the grading guidelines of the exam designers for traditional grading were interchangeable with the checkbox grading schemes: both methods yielded the same score for the same student's solution.

Checkbox grading was developed as an advanced grading method plug-in for Moodle, an open-source e-learning platform (Gamage et al., 2022), months before the experiment.

The exam, including the checkbox grading schemes, can be found in [Appendix E](#).

4.3.2 Assessor training

The grading was distributed among the 3 exam designers and 7 external assessors. In total, we had 7 female assessors and 3 male assessors. They all had, on average, 3.4

years ($SD = 1.4$) experience as an assessor at the Exam Commission. All 10 assessors have experience in teaching mathematics secondary education in Flanders ($M = 29.4$ teaching years / $SD = 8.6$). 5 are still working in upper secondary education (> 16 years old), 3 are retired mathematics teachers, 2 are no longer teaching but were mathematics teachers beforehand. They were, on average, 53.3 years old ($SD = 8.9$).

Two weeks before the exam, all 10 assessors received online training on how to use checkbox grading in Moodle. Some previous tasks of another exam about calculus were turned into checkbox grading to demonstrate how everything works. During the training, the assessors learned to work with the checkbox grading tool through this demo exam. The demo exam continued to be available after the training. Assessors were encouraged to rehearse once more before they started assessing the actual exam to reduce bias due to learning effects. To avoid influencing the assessors, they were not informed about the research questions.

4.3.3 Student examination

A total of 60 students took the exam. All these students were enrolled in a study direction with advanced mathematics in the curriculum.

To answer the research questions linked to this chapter, we selected 30 student exams that all 10 assessors had to grade. Traditionally, the Exam Commission experiences many students who just come to an exam session to have a look at the tasks as preparation for a following session. As these students leave much of the exam blank, their exams are not very interesting for this study because they would artificially increase inter-rater reliability ([RQ 4.2]), as it is straightforward to agree on a blank student's answer by giving 0 points. Therefore, we first scanned all the exams visually (without judging the answers) and put asides those with more than two empty exam tasks. Next, we randomly selected 30 students' exams out of the remaining exams. The mean age of these 30 sampled students was 18.1 years ($SD = 2.2$ / 20 male, 10 female). The mean scores of these 30 students assessed by the 10 assessors under both traditional and checkbox grading can be found in Table 4.1.

4.3.4 Assessment using checkbox grading & Survey based on the Technology Acceptance Model (TAM)

The answers of all 60 students were scanned and entered into Moodle. All assessors had to assess the 30 sampled students for the research study and 3 additional students to get all 60 students graded. They were unaware that of the 33 students they had to assess, all of them were assessing the same 30 students. To answer [RQ 4.2], we randomly selected half of the assessors to grade the even tasks blind and the odd tasks visible; the other half received the opposite treatment.

The assessors had three weeks to complete their assessment work remotely. They were free to choose the order in which they assessed the exam tasks. To measure the time for [RQ 4.1], they always had to tap a start button when grading a task. As such, the time needed to assess a task could be captured by measuring the time between clicking the 'Start'-button and the online form submission. During the training, assessors were told to correct the student's solution to the student's answer without distractions as

soon as they had tapped the 'Start'-button. Since the experiment uses real exams, assessors could always return and change their previous assessment work. The time for these changes was not registered, which is a limitation of the study. Concretely, our time measurements will always slightly underestimate the real time investment (in all conditions).

After their assessment work, they had to fill in a survey to answer [RQ 4.3]. The questionnaire contained two parts; the first part surveyed some personal information (age, teaching/assessor experience), and the second part their view on checkbox grading. The second part contained two times 12 items, measured on a 7-point Likert scale based on the Technology Acceptance Model (TAM) (Davis, 1989; Davis et al., 1989; Venkatesh et al., 2003). The Likert scale ranged from 1 (= totally disagree) to 7 (= totally agree). The items polled teachers' perceived usefulness, perceived ease of use, anxiety, attitude and behavioural intention to use checkbox grading. Assessors had to fill in these items two times: once for blind and once for visible checkbox grading. This part of the survey can be found in Appendix F.

The second part also contained four open questions:

1. What did you like about checkbox grading?
2. What did you dislike about checkbox grading?
3. You assessed some exam tasks without seeing associated and total scores; how did that feel?
4. How did you like the idea that you could dynamically add items to the correction schemes?

Finally, we asked them if they preferred visible or blind checkbox grading.

4.3.5 Re-assessment using traditional grading

To answer [RQ 4.1] and [RQ 4.2], after three months, assessors were asked to regrade all the sampled 30 exams plus the 3 individual exams, but now with the traditional grading scheme of the exam. The three individual exams were deliberately included so that the assessors did not feel that those 30 exams were somehow more special. Three months should generally be enough time to forget most of the details of the previous assessment work (Averell & Heathcote, 2011). Assessors knew beforehand that some grading work was due in March, but they were not informed it consisted of a reassessment using the traditional method. The time monitoring system was the same as with checkbox grading: after tapping the 'Start'-button, they could enter the score obtained by following the grading instructions such as the one shown in Figure 4.3.

4.3.6 Data analysis procedures

All statistical analyses were performed using R version 4.3.

4.3.6.1 Differences in time [RQ 4.1]

The time spent on all tasks was registered under blind/visible checkbox grading and traditional grading. By summing these individual times, we could measure how much

time each assessor spent in the blind versus the visible condition of checkbox grading and compare these times using a paired difference test. To compare the time differences between checkbox versus traditional grading, we calculated the overall time spent in both methods. For checkbox grading, this is the overall time, regardless of the blind or visible condition. If the assumptions of a paired t -test were satisfied, this was chosen as the paired difference test. If not, the non-parametric Wilcoxon rank test was used. The mean times to assess an exam using different methods and conditions were visualised using boxplots.

4.3.6.2 Differences in inter-rater reliability [RQ 4.2]

In order to compare blind versus visible checkbox grading in terms of inter-rater reliability ([RQ 4.2]), a chance-corrected kappa (κ) was calculated for every task, for the whole exam, and separate κ values for each condition. The used kappa statistic is a generalisation of Fleiss' kappa (Moons and Vandervieren, 2023; see Chapter 3). The measure compares the agreement on the items that were checked, weights them according to their associated partial scores and takes dependencies among the checkboxes into account. 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.

Bootstrapping was used with 10,000 bootstrap samples for each task (and the whole exam) to test if the differences in κ values of both conditions were statistically significant ($H_0: \hat{\kappa}_{\text{blind}} - \hat{\kappa}_{\text{visible}} = 0$). As each condition consisted of a different group of assessors (linking to the even/odd treatment), we used an unpaired bootstrap hypothesis test. The bootstrap samples consisted of a random sample with replacement out of the 30 assessed students' solutions. Along with the significance test, we also used 10,000 bootstrap samples for every κ value to determine the bootstrap 95% confidence intervals.

To interpret the agreement level of the different κ values, we used the method of Interval Membership Probability (Gwet, 2012; Vanacore & Pellegrino, 2022) in combination with the benchmark scale of Landis and Koch (1977). This benchmark scale consists of six ranges of values corresponding to as many categories of agreement: Poor, Slight, Fair, Moderate, Substantial and Almost Perfect agreement for coefficient values ranging between -1 and 0, 0 and 0.2, 0.21 and 0.4, 0.41 and 0.6, 0.61 and 0.8, and 0.81 and 1.0, respectively. The Interval Membership Probability (IMP) method returns the range in which the κ statistic belongs with at least 95% confidence.

When comparing the inter-rater reliability between checkbox grading and traditional grading, only the obtained total scores for each task can be compared, as assessors only communicated these in the traditional grading method. Krippendorff's alpha with ratio weights was used as a chance-corrected measure for inter-rater reliability (Krippendorff, 2004). Using ratio weights means the agreement is highest when the obtained scores from both methods are equal, but some agreement is also assigned when the scores differ based on the magnitude of these differences. Only measuring agreement on matching scores would be too strict as it would not differentiate between a grade difference of, for example, 0/4 and 4/4 on the one hand, & 3.5/4 and 4/4 on the other hand. Again, bootstrapping was used with 10,000 bootstrap samples to test if the differences in α values of both methods were statistically significant ($H_0: \hat{\alpha}_{\text{checkbox}} - \hat{\alpha}_{\text{traditional}} = 0$) and to determine 95% confidence intervals.

4.3.6.3 Assessors' usage and views [RQ 4.3]

To answer [RQ 4.3], we categorised the feedback the assessors added to the checkbox grading schemes (Figure 4.2). The main researcher inductively distilled categories from all the added items; next, the three exam designers coded all the added items to these categories independently. An item could be added to multiple categories. An item was coded under a category when at least 2 of the 3 exam designers agreed. For the assessors' views, we reported the outcomes on the different scales of the Technology Acceptance Model (TAM) and their correlations. Moreover, we report the most common answers in four open questions polling what the assessors (dis)liked about blind/visible checkbox grading.

4.4 RESULTS

4.4.1 Differences in time [RQ 4.1]

4.4.1.1 Between blind and visible checkbox grading

As all assessors graded the even or odd tasks blind/visible, the time spent in each condition is usually compared at the assessor's level (= level of analysis) using a paired difference test. However, these results are biased by the allocation procedure to the conditions as the Wilcoxon rank sum test indicated that, in total, assessing the even tasks of all 30 students ($M = 1.97\text{h} / SD = 1.00\text{h}$) took significantly less time than assessing their odd tasks ($M = 2.51\text{h} / SD = 1.13\text{h}$, $W=84$, $p=.009$), regardless from whether they were graded blind or visible. As a result, assessors' blind/visible grading times are not directly comparable as it depends on whether they graded the even or odd tasks blind/visible. A way to wash out this bias is to compare the total time spent grading every student's exam in both conditions, as each of those totals will contain all tasks and all assessors. On this level, an exam took on average 8.13 minutes (=8 min 8 sec / $SD = 4.07$ min) to grade under the blind condition and 8.63 minutes (=8 min 38 sec / $SD = 4.13$ min) under the visible condition (see Figure 4.6). However, a paired t -test indicated that this time difference between blind and visible grading is not significant, $t(29)=-1.34$, $p=.19$. Although it does not matter due to the non-significant result, this test was executed with inflated degrees of freedom by changing the unit of analysis from assessors to students' exams ($df= 29$ instead of 9), rendering the test too liberal. In total, the mean time spent in the blind condition was 2.20 hours (=2h 11 min 50 sec / $SD = 0.95\text{h}$) and 2.29 hours (=2h 17 min 8 seconds / $SD = 1.24\text{h}$) under the visible condition.

4.4.1.2 Between checkbox and traditional grading

The mean total time assessors needed to grade the 30 sampled exams was 4.19 hours (=4h 11 min 35 sec / $SD = 1.97\text{h}$) for checkbox grading and 2.28 hours (=2h 16 min 52 sec / $SD = 1.08$ hours) for traditional grading. A Wilcoxon Signed-Rank test indicated that this difference was statistically significant, $z=-2.8031$, $p=.002$. On the exam level, an exam took on average 8.39 minutes (=8 min 23 sec / $SD = 3.94$ min) to grade with checkbox grading and 4.56 minutes (=4 min 34 sec / $SD = 2.17$ min) to grade with traditional grading (see Figure 4.6).

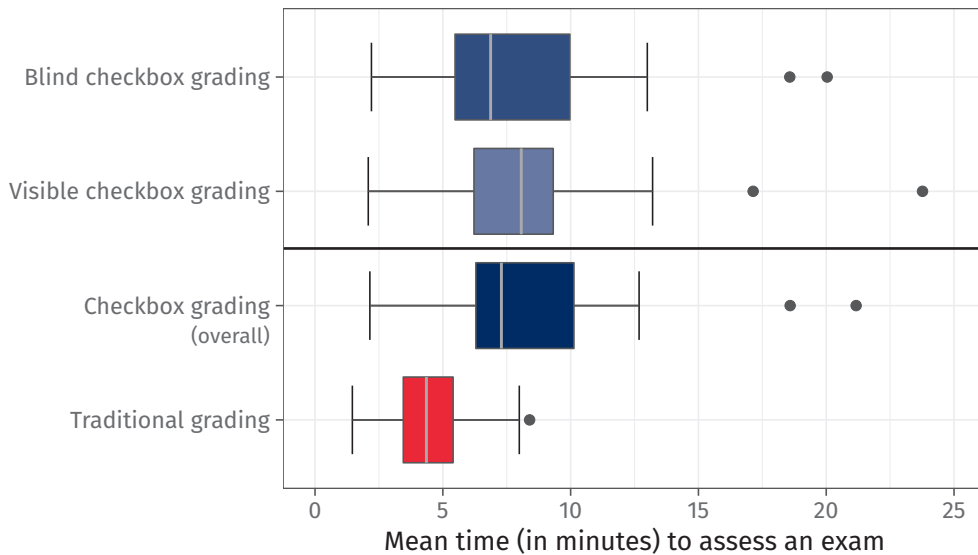


Figure 4.6 – Boxplots of the time (in minutes) to assess an exam under the different conditions

4.4.2 Differences in inter-rater reliability [RQ 4.2]

4.4.2.1 Between blind and visible checkbox grading

The different κ values (Moons and Vandervieren, 2023; see Chapter 3) of all tasks and the whole exam overall and under both conditions can be found in Table 4.2.

Task	Overall		Blind grading		Visible grading		p -value
	κ	95% CI	κ	95% CI	κ	95% CI	
T1	0.803	(0.72 to 0.90)	0.833	(0.75 to 0.94)	0.767	(0.66 to 0.89)	.185
T2	0.641	(0.54 to 0.77)	0.812	(0.72 to 0.92)	0.687	(0.57 to 0.83)	.045*
T3	0.490	(0.40 to 0.61)	0.520	(0.42 to 0.65)	0.420	(0.32 to 0.55)	.007**
T4	0.785	(0.71 to 0.89)	0.723	(0.64 to 0.84)	0.873	(0.79 to 0.97)	.004**
T5	0.835	(0.72 to 0.97)	0.909	(0.81 to 1.00)	0.760	(0.61 to 0.94)	.035*
T6	0.473	(0.15 to 0.88)	0.394	(0.09 to 0.78)	0.586	(0.20 to 1.00)	.052
T7	0.847	(0.72 to 0.98)	0.825	(0.67 to 0.99)	0.892	(0.78 to 1.00)	.343
T8	0.759	(0.65 to 0.90)	0.685	(0.58 to 0.82)	0.652	(0.52 to 0.82)	.564
T9	0.735	(0.65 to 0.84)	0.748	(0.65 to 0.87)	0.733	(0.62 to 0.86)	.828
T10	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)	.224

Table 4.2 – Results of the analysis comparing the inter-reliability of the blind versus the visible condition

The blind condition is significantly more reliable for exam tasks 2, 3 and 5; whereas the visible condition is significantly more reliable for exam task 4. Overall, when calculating the overall κ including all feedback items (weighted according to their score) of the exam, the blind condition has a slightly higher κ value (0.722), than the visible condition (0.698), but this difference is statistically not significant ($p = .224$).

To interpret the agreement level of the different κ values in Table 4.2, the results of the Interval Membership Probability (using the benchmark scale of Landis and Koch, 1977) can be found Table 4.3.

Task	Overall		Blind grading		Visible grading	
	Range	IMP	Range	IMP	Range	IMP
T1	Substantial	0.999	Substantial	0.999	Substantial	0.994
T2	Moderate	0.998	Substantial	0.999	Moderate	0.999
T3	Fair	1.000	Moderate	0.969	Fair	0.999
T4	Substantial	0.999	Substantial	0.975	Substantial	1.000
T5	Substantial	0.996	Perfect	0.960	Substantial	0.953
T6	Slight	0.998	Slight	0.973	Slight	0.956
T7	Substantial	0.999	Substantial	0.999	Substantial	0.994
T8	Substantial	0.981	Moderate	1.000	Moderate	0.997
T9	Substantial	0.994	Substantial	0.989	Substantial	0.977
T10	Substantial	0.999	Substantial	0.994	Perfect	0.952
WHOLE EXAM	Substantial	0.999	Substantial	0.999	Substantial	0.999

Table 4.3 – Magnitude of the inter-rater reliability for all exam tasks under the blind versus visible condition using the method of Interval Membership Probability (IMP)

4.4.2.2 Between checkbox grading and traditional grading

The inter-rater reliability comparison between overall checkbox grading (without distinction between conditions) and traditional grading using Krippendorff's α can be found in Table 4.4. From Table 4.4, it is noteworthy that almost all of the estimates exceed the standard cut-off of 0.8 (Krippendorff, 2004; p. 241–243), indicating a general high agreement on scores between raters in both grading methods.

Regarding the difference in inter-reliability between the two grading methods, both were almost equally reliable; there was only a significant difference for task 3, where assessors agreed better using traditional grading than checkbox grading.

For the sake of completeness, we also compared the inter-rater reliability of the two conditions of checkbox grading with the same groups of assessors in traditional grading (e.g., the group of assessors who checkbox graded the even tasks blind is compared with themselves for the even tasks in traditional grading). As such, it is possible to compare the tasks that were graded using blind or visible checkbox grading with the corresponding tasks in traditional grading. By doing so, it is impossible to compare the

Task	Checkbox grading		Traditional grading		<i>p</i> -value
	α	95% CI	α	95% CI	
T1	0.901	(0.79 to 1.00)	0.894	(0.81 to 0.99)	.893
T2	0.820	(0.70 to 0.95)	0.844	(0.71 to 0.99)	.641
T3	0.695	(0.51 to 0.92)	0.828	(0.70 to 0.98)	.000***
T4	0.900	(0.82 to 0.98)	0.868	(0.77 to 0.97)	0.304
T5	0.953	(0.87 to 1.00)	0.975	(0.93 to 1.00)	0.313
T6	0.578	(0.11 to 1.00)	0.793	(0.21 to 1.00)	0.250
T7	0.965	(0.92 to 1.00)	0.928	(0.83 to 1.00)	0.488
T8	0.751	(0.54 to 1.00)	0.741	(0.54 to 0.97)	0.670
T9	0.854	(0.78 to 0.94)	0.812	(0.71 to 0.93)	0.285
T10	0.920	(0.65 to 1.00)	0.779	(0.47 to 1.00)	0.202
WHOLE EXAM	0.882	(0.76 to 1.00)	0.893	(0.77 to 1.00)	0.245

Table 4.4 – Results of the analysis comparing the inter-reliability of (overall) checkbox grading and traditional grading

inter-rater reliability of the whole exam but only the totals of the even tasks and the odd tasks. The results are shown in [Table 4.5](#).

Task	Blind grading			Visible grading		
	Checkbox α	Traditional α	<i>p</i> -value	Checkbox α	Traditional α	<i>p</i> -value
T1	0.929	0.916	.840	0.883	0.858	.665
T2	0.780	0.881	.083	0.846	0.804	.620
T3	0.642	0.828	.056	0.704	0.821	.205
T4	0.924	0.852	.190	0.867	0.869	.845
T5	0.949	0.949	.690	0.947	1.000	.300
T6	0.466	0.587	.393	0.671	0.867	.722
T7	0.981	0.948	.431	0.950	0.916	.649
T8	0.745	0.717	.423	0.718	0.717	.875
T9	0.873	0.861	.806	0.856	0.744	.089
T10	0.928	0.697	.117	0.900	0.697	.147
Even tasks	0.780	0.775	.902	0.792	0.756	.488
Odd tasks	0.825	0.971	.890	0.868	0.888	.485

Table 4.5 – Results of the analysis comparing the inter-reliability of blind/visible checkbox grading with their equivalents in traditional grading

Table 4.5 shows no significant differences between blind/visible checkbox grading and traditional grading in terms of inter-rater reliability.

4.4.3 Assessors' usage & views [RQ 4.3]

4.4.3.1 Added items to the grading schemes

In the introduction, we explained that assessors could also dynamically add items to grading schemes (see Figure 4.2). 96 items were added throughout the study, which is a lot as all checkbox grading schemes combined initially contained 143 items defined by the exam designers. Their additions could be coded into six broad categories, and the results can be found in Table 4.6. The Fleiss' kappa of the process was 0.72, indicating substantial inter-rater reliability (Landis & Koch, 1977). The table contains double counting, as some items were classified into different categories.

Category	Explanation	Example	No of. occurrences	% occurrences
JUSTIFICATION	Justification for (not) selecting an item	(T7) "Points <i>A</i> and <i>B</i> are marked correctly, but the line is not drawn accurately."	40	39.22%
MATH ERROR	Something was written mathematically incorrect	(T1) " $1 - 3i$ is not between brackets"	30	29.41%
UNCERTAIN	Uncertainty about (not) selecting an item	(T1) "May $\frac{1}{29}$ also be in front of the expression?"	13	12.75%
OBSERVATION	Assessor reports observation, copies a mistake, or indicates good/bad element(s) in the student's answer	(T9) "The student's solution contains only the system of equations belonging to the canonical form of the extended coefficient matrix."	9	8.82%
DISAGREE	Disagree with the scoring system/criteria or a feeling that the score criteria are incomplete	(T2) "No points for the angle 3β ?"	7	6.86%
UNEXPECTED	Student does something unexpected that was not covered by the predefined grading schemes developed by the exam designers	(T2) " z_1 and z_2 were swapped, but calculation is correct."	3	2.94%

Table 4.6 – Categorisation of the added items by the assessors to the grading schemes

Table 4.6 shows that the dynamic item addition system did not effectively capture unexpected student answers, with only three items falling into this category. The

remaining categories primarily contained additions, clarifications, and elaborations on the checkbox grading scheme.

In their answers to the question on how they liked the idea of adding items to the predefined checkbox grading scheme, most assessors indicated they liked the idea; however, they mentioned it was used too often, it was distracting to read all those additions, and they did not always understand the items of their peer assessors. They noticed that many of the items were already covered by the predefined grading scheme developed by the exam designers. One assessor said it was only convenient when the student used a different solution method not covered by the grading scheme.

4.4.3.2 Assessors' views using the TAM model

The assessors' views are reported in [Table 4.7](#). The table contains the mean, standard deviation, Cronbach's α and correlations of the scales stemming from the TAM model. The scales were calculated by averaging the corresponding items responses on a 7-point Likert scale. The corresponding items and their mean and standard deviation can be found in [Appendix F](#). All scales, except for anxiety of visible grading, reached a Cronbach's α higher than 0.7, which is generally accepted as a rule of thumb for scale reliability (Taber, 2018).

Scales	M \pm SD	α	B1	B2	B3	B4	B5	V1	V2	V3	V4
Blind checkbox grading											
B1. Perceived Usefulness	4.66 \pm 1.53	.89	—								
B2. Perceived Ease of Use	4.80 \pm 1.28	.85	.44	—							
B3. Anxiety	3.60 \pm 1.75	.89	-.71	-.75	—						
B4. Attitude Towards Using	4.38 \pm 1.87	.98	.80	.65	-.92	—					
B5. Behavioural Intention to Use	4.43 \pm 1.83	.94	.74	.76	-.93	.96	—				
Visible checkbox grading											
V1. Perceived Usefulness	5.83 \pm 0.71	.81	.54	.21	-.23	.25	.32	—			
V2. Perceived Ease of Use	5.63 \pm 0.93	.82	.53	.66	-.28	.37	.43	.46	—		
V3. Anxiety	2.47 \pm 1.21	.67	-.88	-.40	.68	-.68	-.61	-.42	-.43	—	
V4. Attitude Towards Using	6.08 \pm 0.85	.97	.61	.73	-.63	.66	.75	.69	.66	-.55	—
V5. Behavioural Intention to Use	5.60 \pm 1.28	.87	.60	.74	-.48	.49	.63	.64	.80	-.42	.78

Table 4.7 – Correlation table of assessors' views based on the TAM model

4.4.3.3 What assessors (dis)liked about checkbox grading

In this paragraph, we present a summary of the responses provided by participants to the open-ended questions included in the survey. All assessors acknowledged the value of checkbox grading, specifically the provision of student feedback and the close association between the correct answer and the student's solution. The checkbox grading system was perceived as straightforward and free from grading errors, as they did not have to manually calculate the grade as in traditional grading, as one assessor stated:

"I found it very positive to check the items. In the past, I had to run through the (traditional) grading scheme and wrote down the points for each part. Next, I summed up all the points of the parts. This is not necessary anymore. I even don't need to write down on which parts the student got points; this is now immediately clear for the student and myself!"

Four assessors expressed difficulty in adapting from traditional grading to checkbox grading, citing the perceived rigidity of the new method. Additionally, some assessors felt that the correction scheme initially designed by the exam designers was lacking in appropriate items, and did not view the added items of other assessors as fitting. One assessor conveyed this sentiment as follows:

"Checkbox grading forces you to pigeonhole, and not every pigeon belongs in a hole."

All assessors indicated they preferred visible over blind grading. Only three of the ten assessors indicated in the open question that they had no problem with blind grading. All other assessors mentioned a lack of control by losing the feedback mechanism that points provide:

"I was a bit steerless. We are so used to giving points... and they confirm our assessment."

Others also noted they liked to see the relationship between their assessment, the student's solution and the corresponding grade. Also, fear of missing items to be checked or an alienated feeling when using blind grading was mentioned multiple times:

"Sometimes I don't know if I have checked all the necessary items; when I see the points, I'm sure to have assessed everything. Moreover, assessing without seeing the points felt very impersonal to me."

To conclude, six assessors explicitly mentioned in their answers that the Flemish Exam Commission should adopt checkbox grading as soon as possible. One assessor promptly asked if she could use it in her day-to-day mathematics classroom.

4.5 DISCUSSION

4.5.1 Differences in time [RQ 4.1]

Regarding time investments [RQ 4.1], we see that checkbox grading takes, on average, almost twice the amount of time compared to traditional grading to assess an exam (4.19 hours versus 2.28 hours, see Figure 4.6). The extra time for checkbox grading is consistent with earlier observations that giving written feedback always takes more time than just communicating a grade (Jonsson, 2013). In mathematics education, it is well-known that students' solutions often contain structural error patterns (Movshovitz-Hadar et al., 1987); as such, the same mistakes appear multiple times in different solutions. The same mistakes should lead to the same grade. With traditional grading, assessors may quickly recognise these similar mistakes after a while and assign the corresponding grade immediately. Although the checkbox grading system takes over some of the tasks of the assessors (e.g., calculating the grade and making sure the dependencies in the grading guidelines are enforced), this is not a possibility: assessors repeatedly need to select the same checkboxes in these cases; explaining the increased time investment. It could be argued that this is advantageous because it ensures that each student outcome is considered with sufficient attention to detail and that mistakes are not assumed too quickly. However, this should have resulted in elevated inter-rater reliability for checkbox grading compared to traditional grading, which was not detected (see [RQ 4.2]). A possible mitigation of this problem is letting assessors save and label some combinations of checkboxes. As such, they can immediately opt for that combination. This could be especially convenient when the number of students is high.

Surprisingly, assessors' subjective appreciation of time seems at odds with this objective measurement. One of the questions of the Technology Acceptance Model survey (TAM, see Appendix F) was: 'Checkbox grading allows me to perform my duties as an assessor more quickly.' The assessors largely agreed with a mean of 4.9 out of 7 ($SD = 1.9$) for the blind condition and 5.4 out of 7 ($SD = 1.5$) for the visible condition. Moreover, it was one of the statements they agreed with the most. A possible explanation for this paradox is that the results of the TAM survey showed that assessors thought positively about using (visible) checkbox grading. The benefits of checkbox grading (feedback for students, strict interpretation of the correction scheme) might have made the increased time investment less noticeable. Moreover, it could be that they did not directly compare their time investment with previous traditional grading rounds (which was also not explicitly mentioned in the survey item).

Of blind grading, it was thought that it might take a bit longer, as assessors might dwell longer on selecting checkboxes because they miss the feedback loop provided by seeing the partial and total grades, but this was not the case: no significant time differences could be found between visible and blind checkbox grading.

4.5.2 Differences in inter-rater reliability [RQ 4.2]

4.5.2.1 Between blind and visible checkbox grading

We expected that blind checkbox grading would enhance inter-rater reliability compared to visible grading, as possible conflicts between an assessor's holistic grade and the calculated grade are prevented (Dawson, 2017; Huot, 1990; Moskal & Leydens, 2000;

Stellmack et al., 2009). However, the results reveal a mixed picture in Table 4.2: for the exam as a whole and 6 of the 10 exam tasks, no significant difference was found between the inter-rater reliability of blind and visible grading. Blind grading significantly improved the inter-rater reliability compared to visible grading on 3 exam tasks; visible grading improved the inter-rater reliability on task 4.

A possible explanation for why blind grading outperformed visible grading in terms of inter-rater reliability for tasks 2, 3, and 5 is the strictness of the correction scheme. For example, in task 3, one checkbox could only be selected if a list of predefined keywords was included in the student's answer (see checkbox 'right explanation of C_{11} ' in the exam, see Appendix E) which was a stringent 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 taken into account in task 5 (see Figure 4.7): 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 task by checking the box.

(/1.5) Find a set of parametric equations for the plane $\alpha \leftrightarrow 3x - 2y - 11 = 0$

Set of parametric equations is correct +1.5
Attention: other possible solutions exists (other point and/or other direction vectors)

$\alpha \leftrightarrow$ is missing.

The curly bracket { is missing. -0.5

$k, l \in \mathbb{R}$ is missing. -0.5

$$\alpha \leftrightarrow \left\{ \begin{array}{l} x = \frac{11}{3} + \frac{2}{3}k \\ y = k \\ z = l \end{array} \right. \quad k, l \in \mathbb{R}$$

Figure 4.7 – Checkbox grading scheme of exam task 5

In task 4, visible grading exhibits significantly higher inter-rater reliability. Based on an analysis of the assessors' judgements, this is probably related to the relative complexity of the correction scheme for this task. 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. This is the already mentioned feedback loop that visible grades provide. If assessors obtain a similar score for a similar student answer, they might assume their assessment is correct. In contrast, in the blind condition, they have to run through the complex grading scheme again and again without this feedback, making them more prone to errors.

Concerning the magnitude of the inter-rater reliability, we see in Table 4.3 that 7 of the 10 tasks have a substantial or perfect inter-rater reliability with 95% confidence. No tasks have poor inter-rater reliability. However, tasks 2, 3, and 6 need closer inspection. For task 3, we know that the correction scheme was considered too strict, leading to more free interpretations by the assessors, definitely in the visible condition. The low levels of agreement on tasks 2 and 6 are due to some student answers not fully covered by the correction scheme. In these cases, assessors will match their interpretation with

the ‘spirit’ of the correction scheme, leading to diverging assessments. Capturing these ‘unexpected’ student responses in future correction schemes by the exam designers for these task types is necessary to enhance their inter-rater reliability.

All in all, with our data, the interplay between perceiving grades (when they are visible) and interpreting the student’s solution give rise to the cognitive conflict between the *calculated grade* and the *holistic grade*, but seemingly only when the correction scheme is very strict in what is correct or not. In all other cases, the inter-rater reliability differences are insignificant, and visible grading is preferred for the feedback loop visible partial scores provide. In particular, more complex assessment schemes seem to benefit from this feedback mechanism.

The fact that we did not get a more pronounced picture might also be linked to the two limitations of the study: first, we only had 10 assessors, which is not comparable to large-scale state exams where sometimes all mathematics teachers in a country of some level are involved in the grading process. Second, the Flemish Exam Commission assessors do not know the students they are assessing; as such, they are less prone to all sorts of 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 as these biases might influence the interplay between perceiving grades and interpreting the student’s solution on top of the cognitive conflict between holistic and calculated grades.

4.5.2.2 Between checkbox grading and traditional grading

Both Table 4.4 and Table 4.5 show no differences in inter-rater reliability between checkbox and traditional grading, apart from task 3. The strictness of the correction scheme of task 3 led to more similar outcomes using traditional grading than checkbox grading. However, the most important observation is that most Krippendorff’s α ’s exceed the standard cut-off of 0.8 (Krippendorff, 2004; p. 241–243), meaning that for most tasks, both traditional as checkbox grading exhibit high inter-rater reliability, making it improbable that checkbox grading could ever surpass traditional grading. So, a traditional, well-constructed, transparent grading scheme is equivalent in terms of inter-rater reliability to the checkbox grading approach. The small sample size of 10 assessors grading 30 common exams somewhat limits this statement because the Krippendorff’s α ’s were only based on total scores, thereby losing lots of information we could consider to compare the inter-rater reliability of blind versus visible checkbox grading. However, given the high Krippendorff’s α ’s, this observation would likely hold in larger sample sizes.

4.5.3 Assessors’ usage and views [RQ 4.3]

The coding of the added items by the assessors in Table 4.6 shows that most items are no enrichment of the predefined grading scheme developed by the exam designers. Indeed, the bulk of the added items consisted of justifications for (not) checking an item (39%), noticing that the student was writing something mathematically incorrect (29%), or observing what the student did (13%). Only 3% of the added items addressed unexpected solution methods not covered by the predefined grading schemes. However, the grading scheme’s dynamic nature was designed with these unexpected solution

methods in mind. As the assessors also noted that too many items were added, a solution to this problem might be to foresee a textbox for additional feedback. As such, assessors have a standard place to assign justifications, observations, and uncertainties without cluttering the grading scheme. In addition, some standard scenarios should be accessible from every exam task. These scenarios should behave as an automated flowchart. Scenarios to be included are students writing something mathematically incorrect, students writing too much (but correct), students solving a different task than the one intended (e.g. due to a transcription error), or students using a different solution method than the one required. For example, the scenario where students solve a different task due to a transcription error, could be that assessors have to check if the solved task is as difficult as the intended task. If so, the assessment might continue (with possibly a small penalty to the grade). If not, the answer is considered wrong and is not further assessed. These standard scenarios would probably be an appropriate replacement for the dynamic addition of items by the assessors, which was retrospectively not liked by the exam designers either, as they found that some assessors were using them to 'bend the predefined grading scheme to their liking.'

All assessors preferred visible grading to blind grading. From the TAM model in [Table 4.7](#), we can see that assessors have a strong attitude towards using visible checkbox grading (6.08/7), high perceived usefulness (5.83/7), and low anxiety (2.47/7). Moreover, the high behavioural intention to use the visible checkbox grading (5.60/7) was highly correlated (0.8) with perceived ease of use (5.63/7). Blind checkbox grading was less appreciated on all scales and exhibited a notable increase in anxiety (3.60/7). The open answers revealed that assessors miss the feedback loop stemming from visible grades in the blind condition.

Overall, they highly appreciated the checkbox grading approach. The relatively rigidity could be compensated by foreseeing an additional feedback box and standard scenarios for foreseeable deviations from the grading scheme.

4.6 CONCLUSION

This study investigated the time investment, inter-rater reliability, usage, and views of assessors when utilising checkbox grading for handwritten mathematics tasks on high-stakes exams. Checkbox grading involves assessors selecting pre-defined feedback items that apply to a student's solution, while the adaptive system automatically determines the grade and allows for blind grading, where partial and calculated final grades are not shown. Traditional grading served as a comparison point, consisting of communicating only the grade based on grading criteria. Referring back to the DiaCoM framework (see [Figure 4.4](#); [Loibl et al., 2020](#)), we explored how changing the cues of assessors influences their diagnostic thinking and resulting diagnostic behaviour, our study yielded several key findings.

First, the time investment required for checkbox grading was almost double that of traditional grading. Despite this increase, assessors subjectively experienced the checkbox grading process as highly efficient, likely due to the provision of feedback to students.

In terms of perceiving grading schemes and interpreting students' solutions, both traditional grading and checkbox grading demonstrated high judgement accuracy, as re-

flected by the high inter-rater reliability estimates. There were no significant differences between the two methods, except for one exam task.

When comparing blind checkbox grading with visible checkbox grading, we observed nuanced traces of cognitive conflict arising when calculated grades did not align with holistic grades, particularly when stringent criteria held significant weight in the final grade. However, this interplay between perceiving grades and interpreting student solutions had a minor impact on inter-rater reliability, affecting only three exam tasks. Assessors also indicated that they were not at ease with blind checkbox grading and missed the feedback loop stemming from comparing the final grade with comparable previous students' solutions.

Process indicators like the survey and the added items to the checkbox grading schemes offer ideas to further enhance the checkbox grading approach, like providing assessors with a space to add additional comments as an alternative to the dynamic adding of items and incorporating standard scenarios for unanticipated student solutions. Making shortcuts available where some frequently used combinations of selected checkboxes can be saved, might reduce the time investment when many students are involved. Despite these potential enhancements, assessors exhibited a strong positive attitude towards using visible checkbox grading and found it useful.

While our study was framed in a high-stake mathematics exam at the Flemish Exam Commission, it is important to acknowledge the context-specific limitations of this research. Further investigations in real classroom settings or larger-scale exams could provide valuable insights into the generalisability of our findings. Additionally, exploring students' perceptions of the feedback received through checkbox grading is a planned avenue for future research (see [Chapter 5](#)).

In conclusion, this study provides valuable insights into the usage and effectiveness of checkbox grading for handwritten mathematics tasks. The findings suggest that checkbox grading can be a time-intensive process, but it gives a deep insight in the process of assessment for both the exam designers as well as the students (feedback), which is also acknowledged by the assessors. Both traditional and checkbox grading demonstrate high judgement accuracy, with nuanced indications that blind checkbox grading might slightly reduce the cognitive conflict between holistic and calculated grades. However, visible checkbox grading is preferred by all assessors.

CRedit authorship contribution statement

Filip Moons: Conceptualisation, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Funding acquisition. ● **Ellen Vandervieren:** Supervision, Funding acquisition. ● **Jozef Colpaert:** Supervision, Writing – review & editing.

HIGHLIGHTS

- ✔ While the previous chapter investigated the assessor's perspective on checkbox grading, this chapter switches the perspective to the students. How do they react to the feedback reports stemming from checkbox grading? Can they make sense of it? These are essential considerations when the exam has to be retaken, or transparency on the obtained score is requested.
- ✔ A questionnaire was distributed to all 60 students who had participated in the mathematics exam at the Flemish Exam Commission, providing them with access to sections of their feedback reports. Out of the group of 60 students, 36 took the time to complete the questionnaire, and 4 of this subset agreed to participate in semi-structured interviews.
- ✔ Students preferred traditional grading over checkbox grading when asked to rank feedback types from more to less comprehensible. However, when interviewed using a think-aloud protocol, students were found to interpret 'checkbox grading' feedback more uncomplicated. Moreover, 97% students agreed on the questionnaire that the Flemish Exam Commission should adopt the method.
- ✔ The student's understanding of checkbox grading was high on average and could not be correlated with their exam score, which means that almost all students – both the high- and low-performing ones – could make sense of it.



☑ Chapter 5

CHECKBOX GRADING OF HANDWRITTEN MATHEMATICS EXAMS WITH MULTIPLE ASSESSORS: HOW DO STUDENTS REACT TO THE RESULTING ATOMIC FEEDBACK? A MIXED-METHOD STUDY

PUBLICATIONS

Moons, F., Iannone, P., & Vandervieren, E. (Under review). Checkbox grading of handwritten mathematics exams with multiple assessors: how do students react to the resulting atomic feedback? A mixed-method study. *ZDM - Mathematics Education*.

Moons, F. (2023). Checkbox grading of handwritten mathematics exams with multiple assessors: how do students react to the resulting atomic feedback? A mixed-method study. *Mathematikunterricht mit digitalen Medien und Werkzeugen in Schule und Forschung. Eine Vernetzungstagung.*, 5 & 6 May 2023 in Siegen, Germany.

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)*, 2-5 February 2022 in Bozen-Bolzano, Italy. <https://hal.science/hal-03753446>

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. 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.

5.1 INTRODUCTION

The ultimate test of any feedback intervention is how well students engage with the feedback content (Jonsson & Panadero, 2018). To create a learning environment where '*proactive recipience*' happens, where students take responsibility for engaging with the feedback they receive (Winstone et al., 2017), it is crucial that students first understand their feedback. If students fail to comprehend their feedback, it will not help them improve (Jonsson, 2013).

In this chapter, we explore a feedback intervention that provides what we refer to as 'checkbox grading feedback' to students on a mathematics exam conducted as part of a large-scale, high-stakes exam series developed in collaboration with the Flemish Exam Commission. According to Winstone et al. (2017), *assessment literacy* refers to students' ability to comprehend and utilise the grading process to evaluate their performance. A feedback intervention can support proactive recipience by enabling students to (1) understand the connection between assessment, learning, and expectations, (2) assess their own and others' performance based on specific criteria, (3) grasp the terminology and concepts used in feedback, and (4) become 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 feedback provided also aims to promote transparency so that students perceive the assessment as fair (Bazvand & Rasooli, 2022).

The present study examines how students perceive feedback from checkbox grading, a semi-automated assessment method for handwritten mathematics exam tasks. Paper-and-pencil tasks remain critical in mathematics assessment because, as Hoogland and Tout (2018) warn, digital tasks often focus on lower-order reasoning skills (e.g., procedural thinking). In contrast, handwritten tasks better assess higher-order thinking skills (e.g., problem-solving). Moreover, Lemmo (2021) highlights substantial differences in students' thinking processes when the same task is posed digitally or paper-based, and Bokhove and Drijvers (2010) point out that handwritten tasks allow students to express themselves more freely.

The following paragraph explains the checkbox grading approach. Subsequently, the feedback intervention study is contextualised within a research framework, and the research questions are formulated.

5.1.1 Checkbox grading

5.1.1.1 Idea

Using checkbox grading, exam designers produce a grading scheme for each task consisting of different feedback items written in an atomic way (see below), anticipating the mistakes students may make in the given question. Next, students solve 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 must select the appropriate feedback item ('check the checkboxes') so the same feedback items can be reused repeatedly.

To allocate grades, exam designers can associate items with partial points to be added (green items in Figure 5.1) or subtracted (red items in Figure 5.1). 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 okay or as a signal for the system to know how to proceed). 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 or as a signal 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 grades. 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 sequencing is the main characteristic of the approach. This adaptive grading approach resembles a flow chart that automatically determines the grade, and — ticking the items that are relevant to a student's answer — might at the same have several other benefits: (1) a deep insight into how the grade was obtained for both the student (feedback) as well as the exam commission, and (2) a straightforward way to grade work with multiple assessors.

[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 extended coefficient matrix. (1 point)

Student's answer

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

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

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

$$\stackrel{R_2 - R_1}{=} \left[\begin{array}{ccccc|c} 1 & 1 & 0 & 1 & 2 & -1 \\ 0 & 1 & -4 & 0 & -2 & -4 \end{array} \right]$$

$$\stackrel{R_1 - R_2}{=} \left[\begin{array}{ccccc|c} 1 & 0 & 4 & 1 & 4 & -5 \\ 0 & 1 & -4 & 0 & -2 & -4 \end{array} \right]$$

Solution key

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

$$V = \left\{ \underline{(-5 - 4k - l - 4m)}, \underline{4 + 4k + 2m}, \underline{k}, \underline{l}, \underline{m} \right\}$$

$| k, l, m \in \mathbb{R}$

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

Figure 5.1 – Checkbox grading scheme of exam task 4

An example of this approach is given in Figure 5.1. 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 items 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 can write an additional feedback remark.

Finally, the name ‘checkbox grading’ was inspired by the bestseller ‘The Checklist Manifesto’ (Gawande, 2009), 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.”

5.1.1.2 Link with atomic feedback

The feedback in checkbox grading is called atomic feedback (Moons et al., 2022; see Chapter 1). 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. To write atomic feedback, 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 Figure 5.1). 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 Figure 5.1).

5.2 RESEARCH FRAMEWORK & QUESTIONS

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 Figure 5.5 to compare all these feedback types). The traditional grading scheme of task 2 is shown in Figure 5.2. In the traditional grading process, the assessors receive a PDF file with the grading scheme of all exam tasks, have access to the scanned files of the student’s exams, and only communicate a grade for each task based on these grading schemes. During a review appointment, students receive their exams, the traditional grading schemes and the grades the assessors obtained by applying the

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

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

a) $-5 \cdot z_1$

$-5 = 5 \cdot (\cos 180^\circ + i \cdot \sin 180^\circ) \Rightarrow$
 $-5 \cdot z_1 = 5b \cdot (\cos(\alpha + 180^\circ) + i \cdot \sin(\alpha + 180^\circ))$

OR:

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

- 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 $-5b \cdot (\cos \alpha + i \cdot \sin \alpha)$, 0/1,5 if only $-5b \cdot (\cos \alpha + i \cdot \sin \alpha)$ was written down.

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

b) $z_1 \cdot z_2^3$

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

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

NOTE: 0,5/1 for $b \cdot c^3 \cdot (\cos \alpha + i \cdot \sin \alpha) \cdot (\cos 3\beta + i \cdot \sin 3\beta)$

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

Figure 5.2 – Traditional grading scheme of exam task 2

marking scheme to their exams. In doing so, students sometimes have to guess which criteria were applied to arrive at their particular grade.

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).

In 2016, Lipnevich et al. proposed a student-feedback interaction model 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.

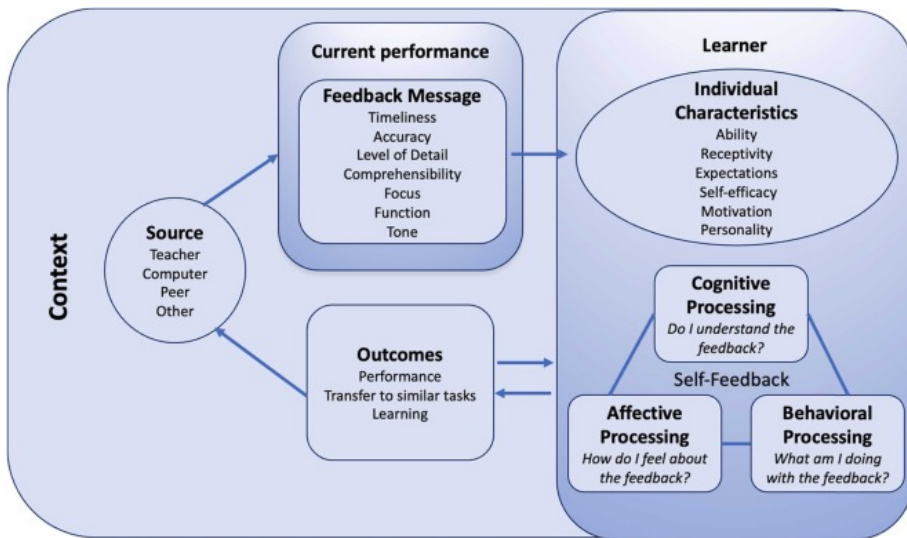


Figure 5.3 – The revised student-feedback interaction model by Lipnevich and Smith (2022)

In our study, 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 solely consists of the Flemish Exam Commission, as students most often do not know each other when participating in such exam. We vary the feedback message with a primary focus on checkbox grading feedback. We gather most *individual characteristics* through a questionnaire, as well as glimpses of the *cognitive processing* 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.

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

- [RQ 5.1] To what extent do students perceive feedback messages generated through checkbox grading as preferred and easier to understand than classical approaches (such as traditional grading, written feedback, or communicating grades)?
- [RQ 5.2] How do students understand (*cognitive processing*) and feel (*affective processing*) about feedback reports from checkbox grading?

5.3 METHODS

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.

5.3.1 Study design

The study was conducted with the Secondary Education Exam Commission of Flanders (the Dutch-speaking part of Belgium). Flanders is a region without any central exams (Bolondi et al., 2019): every secondary school decides autonomously on the assessment of students. Consequently, the 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. A timeline of the study can be found in Figure 5.4.

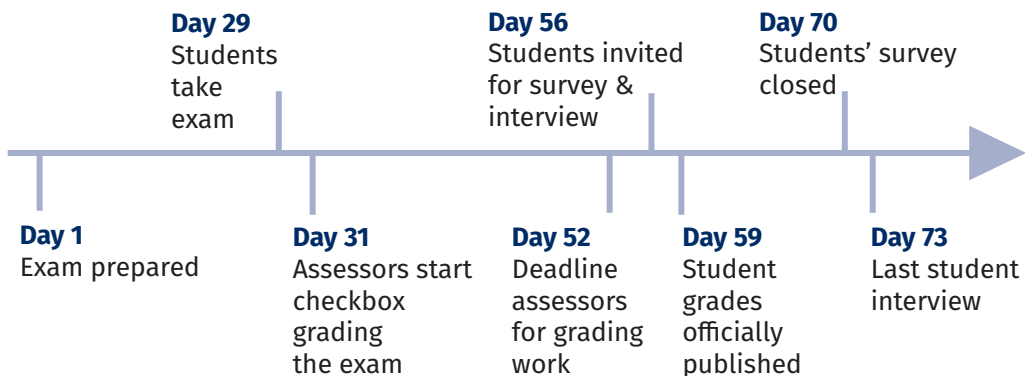


Figure 5.4 – Timeline of the study

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. To pass the advanced mathematics track of the Flemish secondary education senior years (11th/12th grade), the Exam Commission organises two exams that should add up to a passing score above 50%. Although this exam is often called the 'second exam', there is no mandatory order in which to take the two exams, as they cover different topics. The topics for the second exam include complex numbers, matrices, solid geometry, discrete mathematics, statistics, and probability. 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 handwritten tasks is considered. However, as students see the exam as a whole and as we survey, for example, their expected result, they will report their expectations for the whole exam. Therefore, the overview of the exam content in

Table 5.1 also includes the digital part. The paper-and-pencil tasks vary considerably in points allocated based on the importance of the topic in the curriculum; 0.5 points was the smallest partial score.

#	Topic	Learning goal	Max. score	Avg. score M ± SD
Handwritten tasks			54	17.07 ± 9.84
T1	Complex numbers	Calculations with complex numbers in $a + bi$ -form	2.5	1.70 ± 0.97
T2	Complex numbers	Calculations with complex numbers in polar form	2.5	0.59 ± 0.64
T3	Matrices	Modelling with matrices	3.5	1.79 ± 1.11
T4	Matrices	Coefficient matrices of linear equations	3.5	1.19 ± 1.05
T5	Solid geometry	Parameter equations of a plane	1.5	0.13 ± 0.37
T6	Solid geometry	Cartesian equation of a line	1	0.07 ± 0.21
T7	Solid geometry	Drawing a line segment in the x, y, z axis system	2.5	1.16 ± 0.76
T8	Solid geometry	Determining the distance between a point and a line	4.5	0.60 ± 1.30
T9	Solid geometry	Parallel lines in solid geometry	2.5	0.87 ± 0.93
T10	Probability	Modelling a probability experiment	4	0.49 ± 1.12
Digital part			46	20.41 ± 8.52
Algebra			3	1.76 ± 1.12
Solid geometry			6	3.89 ± 2.17
Discrete mathematics			11	2.59 ± 2.30
Statistics			21	10.69 ± 4.59
Research competencies			5	1.48 ± 1.79
Total			100	37.48 ± 16.82

Table 5.1 – Content of the mathematics exam, including the maximum, mean and standard deviation of the scores of the students who filled in the questionnaire

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. All the paper-and-pencil tasks of the exam, including the checkbox grading schemes, can be found in [Appendix E](#).

On the 29th day (see [Figure 5.4](#)), 60 students took the exam. All the students were enrolled in a route with advanced mathematics in the curriculum. Next, the assessors had three weeks to grade the paper-and-pencil tasks using checkbox grading (from day 32 till 52). Checkbox grading was developed as an advanced grading method plug-in for Moodle, an open-source e-learning platform (Gamage et al., 2022) prior to the experiment. The digital part was, of course, assessed fully automatically. The days

between the exam and the start of the assessment were used to scan the answers of all 60 students and input them into the system.

Four days after all assessors finished their work, the questionnaire was ready to be sent to all students. As an incentive, students received personal checkbox grading feedback on three exam tasks during the questionnaire (see [section 5.3.2](#)) and could see their results immediately after completing the questionnaire. If they completed the survey immediately, they would know their results three days before the official date. 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 minutes 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 the traditional review appointment in Brussels.

5.3.2 Questionnaire

5.3.2.1 Instrument development

The questionnaire was implemented in Qualtrics and consisted of four parts and was developed based on the revised student-feedback interaction model by Lipnevich and Smith (2022). A key aspect was to keep the completion time below 15 minutes to motivate students to answer truthfully until the end (Yan et al., 2010). The four parts of the questionnaire were:

1. Individual characteristics & past experiences

The first part gathered some personal information about the students (age, study direction, reasons to get a high school degree through the Exam Commission, number of exam attempts for advanced mathematics, and expected grade). Based on Lipnevich and Smith (2022), we also asked how the students experience the current feedback practices at the Flemish Exam Commission (*context*) and their motivation for mathematics as a school subject. All this information is summarised under the sample description (see Participants, [section 5.3.4](#)).

2. Ranking exercise on the comprehensibility of feedback types

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 5.5](#).

3. Quiz on the understanding of feedback given to a fictional student

In the third part, students saw the feedback report depicted in [Figure 5.1](#). They were asked to answer 10 short false/true questions about the content of the feedback. 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

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

Calculate $\frac{1+3i}{-2-5i}$ and write the answer in $a+bi$ form.

Show all your intermediate steps, don't use your calculator.

$$\begin{aligned} \frac{(1+3i)(-2+5i)}{(-2-5i)(-2+5i)} &= \frac{-2+5i-6i+15i^2}{4-10i+10i-25i^2} \\ &= \frac{-2-15+5i-6i}{4+25} \\ &= \frac{-17-i}{29} = \frac{(-17-i)}{29} \cdot \frac{29}{29} \\ &= -493 - 29i \end{aligned}$$

Solution key

$$\begin{aligned} \frac{1+3i}{-2-5i} &= \frac{1-3i}{-2-5i} \\ &= \frac{(1-3i)(-2+5i)}{(-2-5i)(-2+5i)} \\ &= \frac{-2+5i+6i-15i^2}{4+25} = \frac{13+11i}{29} \\ &= \frac{13}{29} + \frac{11}{29}i \end{aligned}$$

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

Traditional grading

No points when no intermediate steps/solution method provided. Can't be solved by using the polar form of complex number because must be solved without calculator!

$$\frac{1+3i}{-2-5i} = \frac{1-3i}{-2-5i}$$

- 0.5 point (or 0) for correct complex conjugate in the numerator
- NOTE: if not applied or wrong follow through with mistake, max 15/25 because: 0/0.5 for correct complex conjugate in the numerator 0/0.5 for correct final answer (last step)

$$= \frac{(1-3i)(-2+5i)}{(-2-5i)(-2+5i)} \quad (*)$$

- 0.5 point (or 0) for multiplication with the conjugate binomial in the denominator
- NOTE: denominator may be calculated immediately (=29)
- NOTE: (2-5i) is also fine (denominator in this case = -29)
- NOTE: also fine if more steps were used (e.g. first: (2+5i), next: (21+20i))
- NOTE: if binomial conjugate is wrong or missing, no points for the rest of the student's solution

$$\begin{aligned} &= \frac{-2+5i+6i-15i^2}{4+25} = \frac{13+11i}{29} \\ &= \frac{13}{29} + \frac{11}{29}i \end{aligned}$$

- 0.5 point (or 0) for correct calculation of the numerator with intermediate step
- 0.5 point (or 0) for correct denominator (=29 or = -29)
- 0.5 point (or 0) for correct final answer in $a+bi$ form if obtained from (*) with at least 1 intermediate step

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

Checkbox grading

! Checking the calculation

- Correct complex conjugate $1-3i$ in the numerator. +0.5
- If the complex conjugate in the numerator is miscalculated or not applied, the student's answer will deviate from the solution key. Therefore, it is necessary to check the student's calculation individually for the indicated items.
- Check individually: Correctly multiplied by the conjugate binomial in the denominator. +0.5
 - Denominator may also be calculated immediately (= 29)
 - $-(2-5i)$ is also fine (denominator in this case = -29)
 - Also fine if more steps were used; eg. first: $-(2+5i)$, next: $-(21+20i)$
- Check individually: Correct calculation of the numerator with intermediate step +0.5
- Correct denominator (=29 of =-29) +0.5
- Correct final answer in $a+bi$ form +0.5 if calculation is fully correct

Grade: 1/2.5

Only a grade

Grade: 1/2.5

Figure 5.5 – One of the two ranking exercises on the comprehensibility of different types of feedback on the same student's solution of exam task 1

previous questions (as some following questions sometimes revealed the answer of a previous one). As 'understanding the given feedback' can be seen as a latent construct, we analysed the composite reliability (Brunner & Süß, 2005) of the 10 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 can be found in [Table 5.3](#).

4. **Personal checkbox grading feedback: student's cognitive & affective processing**

In the last part, students received a link to access their personal feedback on exam tasks 1, 7 and 10 (see [Appendix E](#)). 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 5.7](#).

5.3.2.2 Analysis

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).

5.3.3 Semi-structured interviews

5.3.3.1 Protocol

The semi-structured interviews of students took place at most a week after they indicated in the survey that they agreed to be interviewed. 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 indicated interesting solutions for exam tasks to discuss. The chosen exam tasks were usually partially correct or incorrect, as these are the best trigger 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?'), and this will be indicated in the discussion of the results. The researcher shared his screen during the online interview to show the students their feedback reports. When traditional grading was discussed, the students saw the traditional grading scheme of the task ([Figure 5.2](#)) and their solution. When checkbox grading was discussed, they saw their complete feedback report like in [Figure 5.1](#). The interview protocol contained two interview questions:

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.




The two questions were inspired by the study conducted by O'Donovan et al. (2004). The authors developed an intervention in which students were provided with marking criteria. These students were then tasked with evaluating and providing feedback on two sample assignments. Next, they discussed their rationales in small groups and with the lecturers. Ultimately, the lecturers provided their assessment of the two tasks. The intervention yielded noteworthy outcomes, as the students exhibited considerable improvement in their performance. By gaining a deeper understanding of how their responses would be evaluated, they could to self-monitor and enhance their work.

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 focus of the study.

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?'

5.3.3.2 Analysis

In the preparatory stage, interviews were transcribed verbatim by the researchers. A straightforward 'traffic light coding' procedure was implemented for each exam task x discussed during the interview (x denotes the number of the exam task):

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

Another researcher double-checked the coding. The results section briefly discusses the students' answers and highlights interesting interpretations with verbatim quotes.

5.3.4 Participants

5.3.4.1 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: 21 students specialise in Sciences & Mathematics, 10 in Economy & Mathematics and 5 in Latin & Mathematics. Reasons to get their secondary education degree through the Exam Commission included: 15 students wanted to graduate faster than possible in a regular high school, 9 felt not at home in a regular school, 5 students faced circumstances that made it impossible to follow regular education (e.g., living

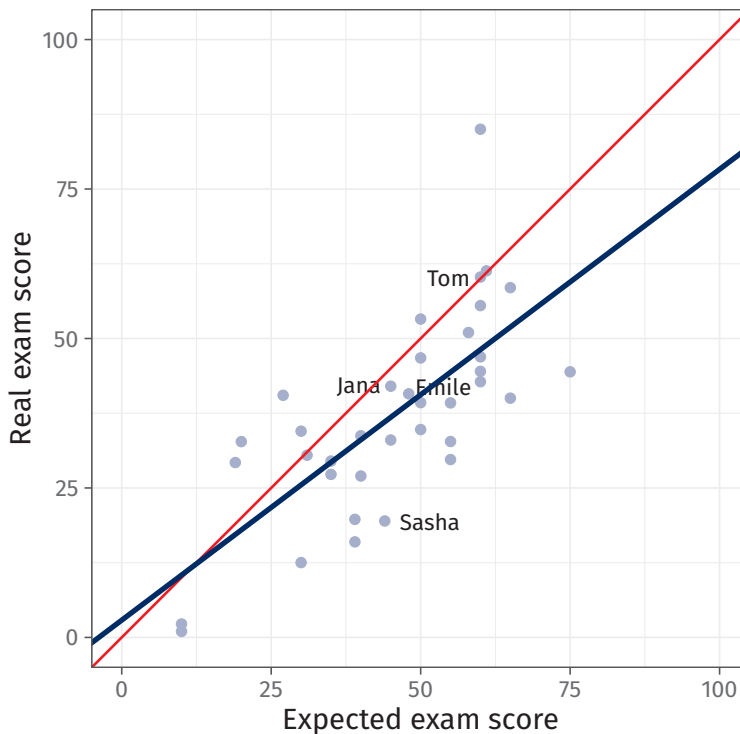


Figure 5.6 – Scatter plot of the expected versus total exam score. The trend line is indicated in blue, and the identity line in red. The students who participated in the interviews are labelled.

abroad, illness, being in an institution), 2 were mature students, and 5 did not provide this information.

The exam results of the 36 students who participated in the questionnaire can be found in [Table 5.1](#)¹. It is noteworthy that the exam results were, on average, relatively low. Indeed, it is a known fact that many students just come to an exam session to know what preparation they need for the following session. Exactly half of our participants took this ‘second’ exam for the first time, 14 for the second time and 6 for the third time.

With regard to motivation, 74,9% of the students say they like mathematics. However, a majority (55.6%) indicate they find it hard to study mathematics on their own (which is the case when taking exams at the Flemish Exam Commission), and 77.8% finds it the most challenging subject at the Exam Commission. 25.0% thinks about switching to a study without advanced mathematics if their exam attempts remain unfruitful.

Concerning the *context* of the student-feedback process, most students (52.8%) already attended a review appointment after a previous mathematics exam. All students said they attended to be better prepared the following time. 57.9% indicated they had problems understanding how grades were obtained, and 84.2% that the assessment

¹This table is different from [Table 4.1](#). In the previous chapter, the results were reported on the handwritten tasks of the 30 exams graded by all assessors. This table provides a complete overview of the results of the students who participated in the survey.

practices are stringent. Regarding the student-feedback interaction model, we know expectations are crucial for feedback receptivity (Eva et al., 2011; Lipnevich & Smith, 2022). Therefore, a simple linear regression was calculated to predict the exam score based on the expected exam score of the students. A significant regression equation was found ($F(1,34) = 36.98, p < .001$), with an R^2 of 0.52. The predicted real exam score equals $2.89 + 0.75 \times (\text{expected exam score})$ when the score is measured in percentages. The scatter plot and trend line can be found in Figure 5.6. As most expected exam scores are below the identity line, most students slightly overestimated their performance.

5.3.4.2 Interviews

Four of the 36 students who filled in the questionnaire agreed to be interviewed:

- **Sasha**, female, 17 years old, studies Economics & Mathematics. She had some negative experiences with the first mathematics exam, where she only received her grades on the review appointment and never saw the traditional grading schemes. She failed this first exam only narrowly. She scored 19% on this second exam but is determined to do better next time.
- **Jana**, female, 17 years old, studies Sciences & Mathematics. Jana wants to graduate quicker from high school than in regular education. She scored 42%. For the moment, she failed both mathematics exams. She wishes that the exam commission would organise more possibilities to take mathematics a year so that she can graduate quicker.
- **Tom**, male, 17 years old, studies Sciences & Mathematics and does not feel at home in a regular school. He had a second and successful attempt on this second exam: he scored 60%, combined with the first exam score, enough to pass advanced mathematics. However, he is still disappointed in his results as he scored only 40% on the first mathematics exam and expected a higher result from this second one.
- **Emile**, male, 19 years old, studies Economics & Mathematics. He failed high school the previous year due to (in his own words) a lack of studying. He retakes mathematics, physics, and English. He passed the first mathematics exam of the advanced track but failed the second part for the second time. He was ill two weeks prior to the exam. He scored 41%, which was just two points short of passing advanced mathematics along with his score for the first exam.

The exam scores of the four students are labelled in the scatter plot in Figure 5.6.

5.4 RESULTS

5.4.1 Survey

The results of the ranking exercise on two different exam tasks can be found in [Table 5.2](#). On average, the traditional grading schemes are preferred above checkbox grading, classic written feedback and only a grade. Only a grade is by far the least preferred option.

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%

Table 5.2 – Results of the ranking exercise on the comprehensibility of feedback types

The quiz results on understanding the feedback report displayed in [Figure 5.1](#) can be found in [Table 5.3](#). On average, the students scored 72% (*SD*: 18.2%). A Pearson correlation coefficient was computed between students' quiz and exam scores. There was no correlation between the two variables, $r(34) = -0.02, p = .91$ with 95% CI $[-0.34, 0.31]$.

#	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	83.3
6	The student gets only 1 point for sub-task b because the quintuples are not enclosed by brackets.	False	50.0
7	If the student had written down a completely correct solution set, the total of this task would have been 2.5/3.5.	True	86.1
Mean ± SD		72.6±18.2	

Table 5.3 – Quiz on the understanding of the feedback shown in [Figure 5.1](#)

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 [Figure 5.7](#). 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.

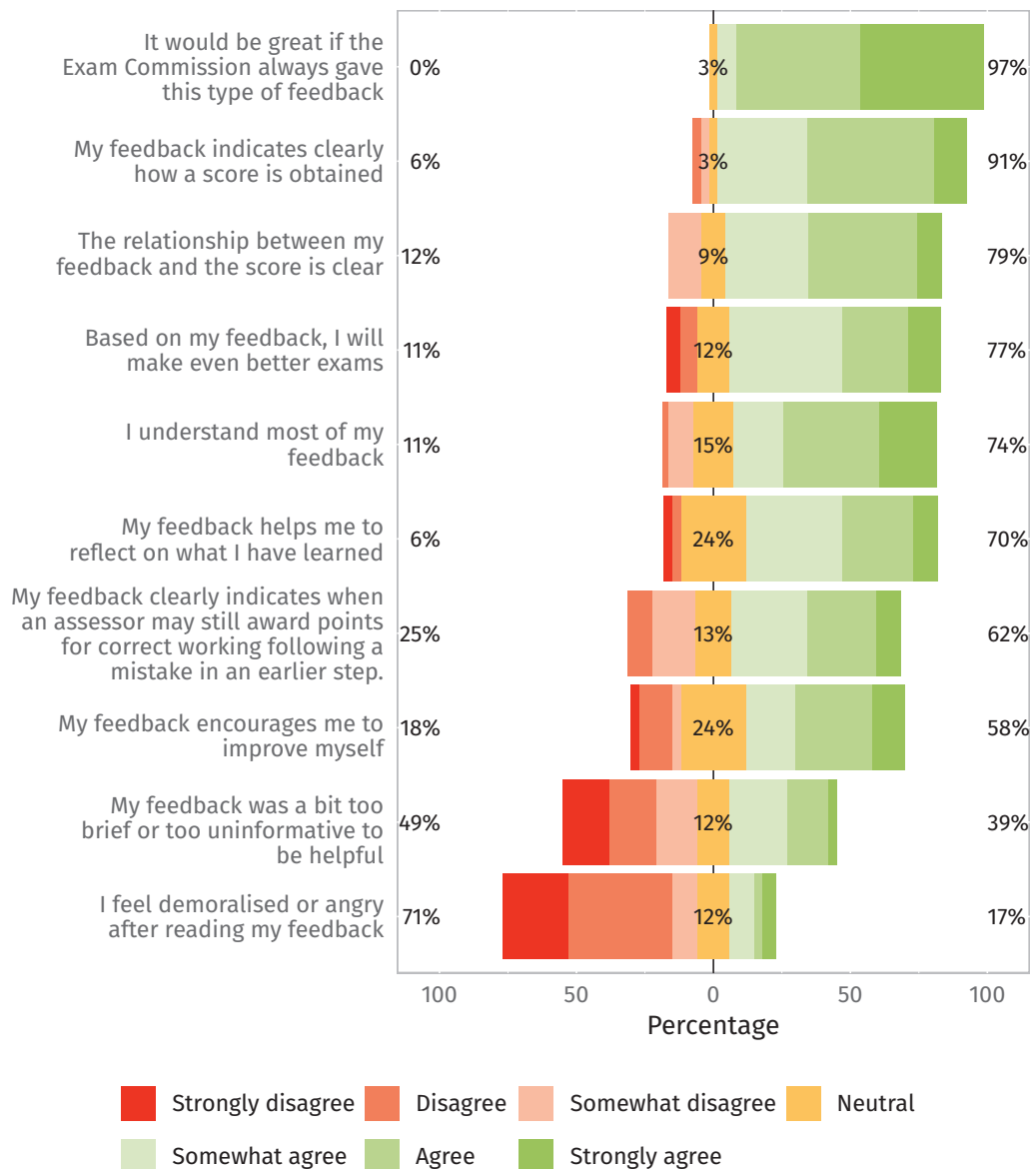


Figure 5.7 – Overview of the students' survey items corresponding to their personal 'checkbox' feedback

5.4.2 Interviews

#	Sasha	Jana	Tom	Emile
Cognitive processing of traditional grading				
1	2	2	2	2
Cognitive processing of checkbox grading				
2	1	3	1	3
3	7	1	4	4
4	10	10	7	1
5	4		8	10
6			9	

Table 5.4 – Results of the ‘traffic light coding’ of the discussed exam tasks

5.4.2.1 Cognitive processing of traditional grading

Table 5.4 presents a rather negative picture of the students’ interpretation of the traditional grading schemes on task 2 (see Figure 5.2). Only Sasha, who answered $-5b \cdot (\cos(\alpha - 5) + i \sin(\alpha - 5))$ on 2(a), independently came to the correct conclusion that 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 $-5[b \cdot (\cos \alpha + i \cdot \sin \alpha)] = -5b(\cos \alpha - 5 + i \sin \alpha - 5)$ 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:

“I don’t really know. Because, well, I wrote α ’s without $+180^\circ$, so instead of $\alpha + 180^\circ$, but they (the grading scheme, ed.) don’t tell anything about this.” (Jana)

Tom, who wrongly answered $-5b(\cos(\alpha) + i \cdot \sin(\alpha))$ 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 $b \cdot e^{i\alpha} \cdot c \cdot e^{i3\beta} = bc(\cos(\alpha + 3\beta) + i \sin(\alpha + 3\beta))$, 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:

“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.

5.4.2.2 Cognitive processing of checkbox grading

The outcome of interpreting the checkbox grading (Table 5.4) is more positive. When seeing their personal checkbox grading scheme, the students could independently draw correct conclusions in 11 of the 16 discussed exam tasks.

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 calculation. At the same time, the instructions indicated that all intermediate steps should be shown as no calculator could be used (see Figure 5.5). 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):

“Oh my God. I realise now that I should have written all my intermediate steps for every task on my exam and not on my scrap paper. It indicates ‘No intermediate steps provided’, which gives a zero because then you might have done it using a calculator. But I did not use a calculator; I could not even work with the one provided!” (Sasha)

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 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 had to change in her solutions when taking the exam a second time. To conclude, Sasha showed a high assessment literacy with checkbox grading that 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 what she missed in sub-task 3(c):

"I had to explain the meaning of element c_{11} (of matrix C , ed.), and from that (the checkbox grading scheme, ed.) I can clearly see that certain keywords needed to be in my answer, and I didn't mention them; therefore, I didn't get any points for that. For the explanation of c_{12} , I don't see any keywords because the most important thing is that the element has no meaning, and I wrote 'This means nothing', which more or less equals the solution key. Well, the wording is not exactly the same, but it does mean the same thing." (Jana)

She repeatedly stressed how clear she found the checkbox grading scheme, for example, when discussing task 1 (see [Figure 5.5](#)):

"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 any more 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 interview started with his correct solution to task 1 (2.5/2.5). On the hypothetical question 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 5.4 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 to understand; therefore, this task was coded in red in Table 5.4. Interestingly, this was only the first time during the interviews that a lack of content knowledge is the cause of a lack of understanding of the feedback.

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?! Ooooooh, I have written 0.013 and not 1.013. I had to add 1! Ooooooh nooooooo. Oooh, such a bummer! That would have been that 2%! That would have been that 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.

5.5 DISCUSSION

Returning to our two research questions, the first on the preference of feedback messages and ease of interpretation; and the second on the cognitive and affective processing of checkbox grading, we now synthesise the results.

5.5.1 Preference & ease of understanding of different types of feedback messages [RQ 5.1]

Regarding preference, the traditional grading schemes of the Flemish Exam Commission were, in general, chosen above checkbox grading, classic written feedback, or only

communicating a grade. It is hardly surprising that only communicating a grade takes the last place, but the position of classic written feedback is interesting. These contained a few to-the-point, personalised sentences explaining what the student did wrong or was missing. Students should only make an effort to understand these sentences. Moreover, these sentences are straightforward and self-explanatory and do not require students to relate their solutions to the grading guidelines or solution key. Remarkably, this approach is less preferred than the two other approaches (checkbox and traditional grading), which are both a form of grading criteria. This finding is in line with the study 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 whilst evaluating their learning processes and outcomes.” (p. 283). Another potential explanation is students’ limited familiarity with receiving classic written feedback in mathematics. Studies such as Knight (2003) have shown that students in mathematics classes typically receive grades for their tests, with only occasional instances of mistakes being highlighted. However, explanations accompanying these grades are seldom provided. It could also be that students missed a clear link between the grades on the intermediate steps in the classic written feedback, as traditional and checkbox grading show how a grade is constructed.

A similar mechanism might have occurred when comparing traditional grading with checkbox grading. Students saw checkbox grading reports for the first time when ranking their preference, so possibly they opted for what they were familiar with. Indeed, the first place for traditional grading somewhat contradicts the fact that 57.9% of students who already attended a review appointment (where they have to handle the traditional grading schemes) indicated that they needed help to link their grades with the grading criteria. At the same time, 91% agree that checkbox grading clearly indicates how a score is obtained.

In the student interviews, only one student (Sasha) could independently infer the grade she received when she compared her solution to the traditional grading scheme, which was less than checkbox grading. Notably, the four students interviewed expressed their liking of checkbox grading without the researcher asking for it.

5.5.2 Cognitive and affective processing of checkbox grading [RQ 5.2]

The results support the hypothesis that checkbox grading helps students understand the assessment and their exam task grades. The results on the quiz (M : 72%) were high compared to the exam scores (Table 5.1), eliciting a satisfactory understanding of checkbox grading. The results were higher than the exam designers of the Flemish Exam Commission and the researcher had anticipated, as the quiz items were rather challenging. Moreover, we could not find a correlation between the quiz and exam scores. If we take the relatively low exam scores into account, this means that most students, even the lower-performing ones, could make sense of the checkbox grading feedback. Also, the questionnaire items of part 4, concerning the personal checkbox grading feedback received on three exam questions, showed appreciation for checkbox grading. 97% indicated that the Flemish Exam Commission should adopt it, 91% understood based on the feedback how the score was obtained, 74% acknowledged their understanding of the feedback.

The ease of understanding, even for lower-performing students, and the appreciation for checkbox grading were also confirmed during the interviews. All students could interpret their checkbox grading feedback on most exam tasks. Three kinds of interpretation problems emerged. First, some emotional responses to the checkbox grading feedback were observed (Sasha on task 1, Tom on tasks 7/8). That emotion may get in the way of feedback understanding is well-known (Goetz et al., 2018; Koenka et al., 2019; Lipnevich & Smith, 2022), making it somewhat likely that their failure to understand would have emerged in other feedback types too. A lack of content knowledge to interpret the feedback (Tom on task 9, Emile on task 4) was the second interpretation problem. However, it is surprising that patchy content knowledge was only a problem twice, given that three of the four interviewed students failed their exams. The last interpretation problem occurred with Tom on task 8, who disagreed with the grading criteria. A limitation of the results of the interviews is the small sample size: with four interviews focusing on different exam tasks, saturation could not be achieved (Hennink & Kaiser, 2022).

When combining the questionnaire and interview results, the main finding is that checkbox grading was understood by most students, even those who scored poorly. There was no correlation between the (high) quiz scores and the (low) exam scores, and only in very few instances, we observed that the lack of content knowledge got in the way of understanding the checkbox grading feedback.

5.6 CONCLUSION

This study aimed to investigate students' perspectives on checkbox grading. This study has identified that in the context of high-stakes mathematics exams, students preferred feedback in the form of grading criteria over classic written feedback or only communicating a grade. When it comes to checkbox grading, students perceive the feedback method positively. It is encouraging that both low and high-performing students demonstrate proficiency in understanding the provided checkbox grading. The small sample size for the questionnaire and the semi-structured interviews limits the study. Nevertheless, it gives a first glimpse of the positive students' perceptions and ease of understanding of checkbox grading.

CRedit authorship contribution statement

Filip Moons: Conceptualisation, Methodology, Formal analysis, Investigation, Writing – original draft, Funding acquisition. ● **Paola Iannone:** Methodology, Writing – review & editing. ● **Ellen Vandervieren:** Supervision, Funding acquisition.

GENERAL CONCLUSION

The general conclusion consists of two parts: first, we revisit the research goals stated in the general introduction; next, we discuss some ideas for follow-up research.

MAIN FINDINGS

In the following paragraphs, we revisit the overarching research goals from the general introduction and summarise and discuss the primary outcomes and implications. The graphical overview of this dissertation — that was depicted in the introduction in [Figure 6 \(page 19\)](#) — drives this general discussion as it links the research questions of the individual chapters to the overarching research goals. As we now glue Part I and II together, we will sometimes refer to them as Study I and II and slightly loosen the distinction between teachers (Part I) and assessors (Part II).

[RG 1] Software development of semi-automated approaches

The first research goal about the software development of semi-automated approaches was foundational to the dissertation thesis but never directly investigated as a research question. In this part, we shortly reflect on this development process and give ways to valorise the developed applications.

Moodle as a development framework for educational research

Both the SA-tool from Part 1 and checkbox grading from Part 2 were developed as an *advanced grading method* in Moodle (Moodle, 2022). At the beginning of the PhD project, there was some doubt about developing the first SA tool from scratch (as a stand-alone tool) or as a plug-in for an existing Learning Management System (LMS). We opted for Moodle, an open-source e-learning platform used by 150 million people worldwide (Gamage et al., 2022). A Moodle installation for this PhD project was installed on the servers of the University of Antwerp. Looking back, we are delighted to have gone for this option. By developing the approaches as advanced grading methods, we only had to focus on the desired functionalities of the SA approaches; all other technical requirements were already available in Moodle (like a user management system for teachers/students, a grade book, an environment to assess students' work, ways of presenting the assessment work). Without using such an LMS, developing, investigating,

and writing up two semi-automated approaches over 3.5 years would probably never have been possible.

Nevertheless, we must also be critical: Moodle is developed in PHP in an object-oriented way, but the Moodle developer documentation (Moodle, 2023) sometimes lacks comprehensive specifications of the objects and methods, especially when programming specialised plug-ins such as advanced grading methods. These require a deep understanding of the software architecture behind the grading system of Moodle, and sometimes the technical specifications are too brief to get a clear insight. Although we could figure out everything in the end, and the development process was much faster using Moodle than programming something from scratch, these were time-consuming hurdles.

Valorisation of the developed software applications

An FWO Strategic Basic Research fellowship funded this doctoral project. One of the goals of these fellowships is that the research results can be valorised, which was an additional argument for choosing Moodle: the developed plug-ins integrating semi-automated feedback and assessment in the open-source e-learning platform Moodle can, in a later phase, be released as an open source project to the public domain. During the PhD, we had talks with Televic, the Belgian company behind assessmentQ. AssessmentQ is a web-based platform that enables organisations to create, manage, deliver and track all sorts of online assessments, also used by the Flemish Exam Commission. The talks discussed the possibilities of integrating checkbox grading into assessmentQ. However, publishing research articles with the ideas behind our semi-automated approaches and, at the same time, economically valorising the results seem not to be compatible. Therefore, it is essential to contribute our developed technology to society by making it available as open-source software after the PhD.

[RG 2] Time-savings

Both [Chapter 1](#) and [Chapter 4](#) investigated how the newly developed semi-automated approaches influenced the time investment of the teacher (assessors) compared to traditional paper-and-pencil approaches (see [\[RQ 1.1\]](#) and [\[RQ 4.1\]](#)). Time investment is an important consideration in feedback practices, not only because having too much assessment work is frequently a complaint about the teaching profession (Eurydice, 2021; Gibson et al., 2015; Jonsson, 2013), but also because it has a direct influence on the content of the feedback. Indeed, Price et al. (2010) showed that one of the coping mechanisms to overcome the workload of giving feedback is shortening feedback. This coping mechanism was also detected in our data, where we saw that classic PP feedback given by teachers tends to be shorter ([Chapter 1](#)) and contains more abbreviations ([Chapter 2](#)) compared to SA feedback.

One of the aims of our research project was to investigate whether semi-automated approaches could alleviate the time demands placed on teachers while maintaining the quality and comprehensiveness of feedback. However, both the study from [Chapter 1](#) (SA tool used by teachers) as [Chapter 4](#) (checkbox grading) did not necessarily result in time savings as initially aspired.

In the first study, the teachers in the sample tended to give significantly more feedback when using the SA tool rather than completing the correction job faster. The study serves as a reminder that claims of time savings should be compared to a reference point and that over time, teachers may become routinised at giving feedback anyway. The study also suggests that some teachers may view attempts to reduce the demands of the teaching profession as opportunities to do even more work.

In the second study, checkbox grading took almost twice the time compared to traditional grading, possibly due to the need to repeatedly select the same checkboxes for solutions containing similar mistakes. Surprisingly, assessors subjectively appreciated the use of checkbox grading as very efficient. Moreover, there was no significant difference in time between visible and blind checkbox grading. For the exam designers, the difference in time investment for developing a traditional grading scheme compared to a checkbox grading scheme was not investigated as the researchers developed these grading schemes to establish the new approach. Moreover, there was no editing tool to develop these checkbox grading schemes; they were directly inserted into the database. Still, as an educated guess, we suspect that developing checkbox grading schemes takes without experience probably a lot more time than writing out traditional grading schemes. Indeed, initially, writing the atomic feedback items, defining the hierarchical dependencies between the checkboxes, and colouring the parts in the solution key can be a Herculean task. However, if a comprehensive editor is accessible to facilitate the reuse of previous exam questions and take care of rendering and layout, the time invested in developing checkbox grading schemes could become comparable to that of creating traditional grading schemes. As exam designers gain more experience, this might be the case. Indeed, it should be noted that drafting and laying out traditional grading schemes is by no means an effortless task either.

Overall, both studies highlight the importance of careful consideration of the potential impacts of introducing technology or alternative assessment methods on time investment, as there seems to be no such thing as a free lunch: feedback, in whatever form, will always be more time-consuming than just communicating a grade or highlighting a mistake (Knight, 2003). Studies on the use of audio feedback (i.e., voicing and recording feedback instead of writing it out) came to a similar conclusion as ours: audio feedback provides a way to increase the amount of feedback communicated to the students as compared with written feedback, but again, did not make the feedback process faster (Huang, 2000; Kirschner et al., 1991; Pearce & Ackley, 1995).

[RG 3] Grading reliability

Grading reliability was a key research question in the checkbox grading study, described in [Chapter 4](#) in collaboration with the Flemish Exam Commission. As they work with several external assessors, guaranteeing a high inter-rater reliability is a crucial component of the high standards of their assessment work. It led to the discovery of a new chance-corrected measure for inter-rater reliability discussed in [Chapter 3](#). More specifically, the research question on the inter-rater reliability of blind versus visible checkbox grading led to the discovery of chance-corrected κ statistic allowing multiple raters to classify subjects into one-or-more categories. The measure is a generalisation of the Fleiss' kappa and was derived in detail in [Chapter 3](#). Next, the measure was used

in [Chapter 4](#) to investigate the differences in inter-rater reliability between blind and visible checkbox grading using bootstrapping.

The results revealed a mixed picture: for the exam as a whole and most exam tasks, no significant difference was found between the inter-rater reliability of blind and visible grading. Blind grading significantly improved the inter-rater reliability compared to visible grading on three questions, and visible grading significantly improved the inter-rater reliability on one question. Referring back to the DiaCom framework (Loibl et al., 2020), it seems that the *diagnostic behaviour* from the assessors is influenced by their *diagnostic thinking*; more specifically, their *interpretation* of the value of partial scores seems to be a vital factor on how the checkbox items are *perceived*, which then influences their *decision making* (too harsh/lenient). Indeed, the assessors indicated that knowing the partial scores connected to a checkbox item is important information to perceive the item. When an item left little room for interpretation but weighed strongly in the final outcome (due to the linked partial score or the sequentiality), assessors diverged more in their assessments in visible checkbox grading: solutions they evaluated as ‘good enough’ were slightly more likely to receive an assessment not fully aligned with the checkbox grading criteria. When the grading scheme has a rather complex sequentiality, grades offer a critical feedback loop for assessors. By seeing the final grade, they can compare it with the grades given to previous students with comparable solutions. This way, they get an indication of whether their assessment can be correct, possibly explaining the significant improvement in inter-rater reliability in visible grading on one exam task. The sample size somewhat limits these results: it consists of 10 assessors who all assessed the same 30 exams. As the large-sample variance of the proposed κ statistic from [Chapter 3](#) still needs to be determined, performing a statistical power analysis of the statistical test to compare the resulting κ 's of blind versus visible checkbox grading is not yet possible. Although the Flemish Exam Commission has no more than ten assessors for mathematics, more extensive studies in different contexts may find more significant impacts of using blind versus visible checkbox grading.

When comparing traditional grading with checkbox grading, we could not find significant differences in grading reliability except for one exam task. So, a well-constructed, transparent grading scheme is equivalent in terms of inter-rater reliability to the checkbox grading approach. Moreover, most Krippendorff α 's had a value near 0.8, indicating high agreement in both grading methods (Krippendorff, 2004). The small sample of 10 assessors grading the same 30 exams is again a limitation, definitely because only the obtained total scores could be compared for each task and the whole exam, making the bootstrapped hypothesis test based on much less information than the one to compare blind/visible grading. However, this observation would likely hold even with larger sample sizes, as the observed high values of the Krippendorff's α for both grading methods make it improbable that one of the methods can surpass the other as further improvement in inter-rater reliability based on scores seems difficult to achieve.

Grading reliability was not analysed in the first part in the end, but the data to do so were collected nonetheless. The intention was to compare how consistent teachers were with themselves when assigning grading in the SA and PP condition by letting them remotely reassess all 60 tasks three months after the experiment, the so-called *intra-rater reliability*. However, more urgent follow-up research questions emerged during the crossover experiment (see [A more intelligent suggestion system](#) and [The adder/subtractor problem](#) under ‘Further research’). These research questions need

to be sorted out first, as they are very likely to impact the grading reliability of the current implementation of the first semi-automated assessment approach. Indeed, the lack of a smart suggestion system to propose feedback items to reuse led teachers to sometimes rewrite feedback items and their associated partial scores because they could not find the pre-existing item they were looking for. The failure of the grade calculation algorithm to flawlessly come up with the grade the teachers intended (see [The adder/subtractor problem](#)) almost certainly further compromises the intra-rater reliability.

[RG 4] Feedback characteristics, content & quality

The fourth research goal was investigated using different methodologies in both parts. In the first part, the concept of atomic feedback was devised and experimentally verified if it enhanced feedback reusability in [Chapter 1](#). Next, in [Chapter 2](#), the feedback from both conditions (SA/PP) was compared and contrasted using text-mining (Ferreira-Mello et al., 2019) and classifying the feedback reports resulting from both SA and PP using a codebook from the literature (Busch et al., 2015a, 2015b). In the second part, in [Chapter 5](#), the student-feedback interaction model (Lipnevich & Smith, 2022) was used to investigate students' cognitive and affective processing of the feedback resulting from checkbox grading.

Atomic feedback

Atomic feedback is a collection of form requirements for written feedback which ran like a thread throughout this thesis. To write an atomic feedback item, teachers must: identify the independent error occurring and write short feedback sentences for each error, independently of each other. In Part 1, we found that atomic feedback could be distinguished from non-atomic feedback within the context of a linear equation task after four iterations. It was confirmed that atomic feedback increases the reusability of feedback items. Moreover, as 73.7% of all feedback items could be classified as atomic after only a short training where the definition of atomic feedback and an example were shared, we could conclude that formulating atomic feedback is easy for teachers to learn. One of the most common violations of atomicity was providing both the location and error in the same feedback item, which could be solved by allowing teachers to tap on the mistake and create feedback items at the location of an error (see [Future research](#)). The second most common violation was addressing multiple mistakes within the same feedback item.

In the second study, we added a requirement to the definition of atomic feedback: 'a knowledgeable assessor must be able to determine unambiguously whether an item applies to a student's answer', transforming feedback items to yes/no-questions which guide an assessor as a flowchart through the grading scheming.

SA vs PP feedback

In part 1, the SA tool with reusable feedback leads to teachers giving significantly more feedback ($d = 0.41$). When comparing feedback given using the SA tool and given the classic PP way, both feedback types had similarities in terms of word usage and frequency, the amount of feedback given on bad, moderate and good solutions, and the

predominance of similar sentiments in feedback reports. Both types of feedback also had an equal frequency of corrective and descriptive feedback as diagnostic activity, as well as giving hints, pointing out misconceptions and missing parts, and writing erroneous feedback. Differences between the feedback types included SA being more comprehensive due to reusability and less use of abbreviations. In contrast, PP feedback was more focused on the main issues, concrete, tailored to the student's solution and analysed the student's solution more often.

From a quality point of view, we found that almost 1 out of 20 feedback reports are erroneous in both conditions. Moreover, the predominance of descriptive and corrective feedback is worrying in cases where the student's solution is well analysable. Some feedback reports seemed 'incomplete': it interpreted the start of a student's solution and stopped when finding a mistake, sometimes missing the overall picture of a student's solution. The incompleteness led to an overall low number of deficits and strengths addressed in both conditions, particularly for PP feedback. On the other side, SA feedback was sometimes less to the point, too, by focusing on all kinds of mistakes, making no distinction between major and minor issues. However, the feedback quality seems more compromised in the SA condition, as fewer teachers gave feedback analysing the student's solution compared to the PP condition. All in all, teachers should not confuse the handiness of semi-automated feedback with quality: a tool to reuse feedback can help them greatly in providing more elaborate feedback to students, but continued attention to feedback quality and pedagogical content knowledge remains critical in all kinds of feedback methods where teacher provide feedback (Busch et al., 2015a, 2015b; Depaepe et al., 2013).

From a methodological point of view, Chapter 2 also highlighted the opportunities and constraints of using text mining for education (Ferreira-Mello et al., 2019) and confirmed the importance of using qualitative research methods to make statements about feedback content and quality (Yu et al., 2011).

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 the pre-defined atomic feedback (developed by the exam designers) to the student giving a detailed insight into what went wrong and how the grade was obtained. Researching the content and quality of the resulting checkbox grading schemes was not an objective in the second part, as it consisted only of a first try of the new semi-automated approach with the checkbox grading schemes being developed by ourselves (based on the traditional grading schemes the Exam Commission provided). However, some differences in the atomic feedback between the two studies are noteworthy.

Differences in the atomic feedback characteristics between the two studies

In the first study, individual teachers built their database with feedback items while providing feedback; in the second study, the checkbox grading schemes were developed in advance. The exam designers created the questions and accompanying grading

schemes based on the requirement of the Flemish curriculum of advanced mathematics. In the developed grading schemes, they anticipate almost any mistake that could occur. This results in a different kind of atomic feedback than in the first study.

Differences like the inclusion of the solution key and the use of colours to connect the feedback item to the right location in the solution key are obvious (see [Appendix E](#)). Still, more subtle differences are noteworthy, too. When scrolling through the checkbox grading schemes, one will notice that most feedback items are phrased positively (e.g. '-5 is correctly converted to polar form'), contrasting the first study's feedback items, which mostly point to errors; this positive formulation in checkbox grading results from the additional requirement that a knowledgeable assessor must be able to determine unambiguously whether an item applies to a student's answer. Indeed, by phrasing most items in terms of being correct, an assessor just has to compare that part of the student's solution with the corresponding part in the solution key. It is much easier to check if a part of the student's solution aligns with the solution key than to check if a particular mistake happened. Moreover, due to previous checkings, the system indicates whether the assessors can make a direct comparison or have to check individually because of a previous mistake. More importantly, by phrasing the checkbox items positively, most mistakes are anticipated because they usually do not result in something correct. The level of detail of the checkbox grading scheme is tightly related to its atomicity: the more it anticipates independent errors, the more students will get partial scores for small steps that were good in their solutions.

Concerning the compliance of the checkbox grading items with the definition of atomic feedback, the exam designers sometimes deliberately violated the independence requirement of atomic feedback. For example, in task 2, one of the items indicates 'Modulus and argument are correct.' In the spirit of atomic feedback, these should be two items: 'Modulus is correct' and 'Argument is correct', which would return a more granulated and thus more detailed grading scheme. Of course, it is the right of the Exam Commission to follow the definition of atomic feedback pragmatically while also considering the difficulty level they want to achieve with an exam task. Lastly, the list hierarchy plays an additional role in checkbox grading. The idea of clustering items that belong together still holds in the second study. However, the list hierarchy is also a guide through the grading scheme: it indicates a dependency between the feedback items: a sub-item can only be selected if the parent item was.

[RG 5] Teachers', assessors' & students' views

The results of the first part showed that teachers had a positive attitude towards using the semi-automated system, perceived it to be useful, and had a strong intention to use it. However, they rated the perceived ease of use lower than all other measures, possibly due to the non-intelligent suggestion system, which made it difficult to find the right feedback item to reuse (see [A more intelligent suggestion system](#) under 'Further research'). Some teachers suggested that preparing feedback items beforehand could be helpful, and others wanted to share their feedback items with colleagues. Surprisingly, these two suggestions were effectuated in the second study.

In the second part, all assessors preferred visible checkbox grading over blind grading and had a stronger intention to use visible checkbox grading, which was highly correlated with perceived ease of use. Blind checkbox grading was less appreciated and increased

anxiety, with assessors missing the feedback loop provided by visible grades in the blind condition.

The second part also investigated the student's views on the received checkbox grading feedback by conducting a survey and semi-structured interviews by the students. Students' perceptions of checkbox grading were compared to other feedback types for high-stakes mathematics exams. We found that traditional grading schemes were generally preferred over checkbox grading, classic written feedback, or only communicating a grade. However, when interviewed using a think-aloud protocol, students were found to interpret 'checkbox grading' feedback more easily. Moreover, 97% students agreed on the questionnaire that the Flemish Exam Commission should adopt the method. The clarity the feedback offers into how their scores were obtained was regarded highly. The student's understanding was high on average and could not be correlated with their exam score. The results suggest that checkbox grading is an effective feedback method for helping students understand their exam performance. Even low-performing students demonstrated proficiency in understanding the provided checkbox grading.

FURTHER RESEARCH

A more intelligent suggestion system

In the experiment described in [Chapter 1](#), many teachers acknowledged they sometimes forget how they had phrased feedback items using the SA tool. As such, they could not find the feedback item they needed, although they knew they had already written a fitting item. This caused them to formulate already given feedback again instead of reusing feedback, which was confirmed by the identification of nearly identical feedback items in their databases. This was due to the non-intelligent suggestion system: it only literally matched what teachers were typing with their items in the database. Improving the suggestion system by incorporating ideas from the extensive literature on *recommender systems* (Mohanty et al., 2020) is a priority for further research. It is a vital gap to make the semi-automated approach described in the first part valuable and adopted by teachers. It is also a necessary step before making the software available. To feed such a recommender system with information on which feedback items might be appropriate, one solution is to allow teachers to indicate where an error has occurred in a handwritten solution, which provides helpful information about which feedback items are appropriate. Additionally, patterns of the use of feedback items can be unravelled (e.g. which items are popular, which often co-occur together) and added to the suggestion system's predictive ability to suggest appropriate items.

The adder/subtractor problem

When inspecting the graphical dissertation outline in [Figure 6](#), a blind spot of the research project emerges: grading reliability was not yet investigated for the SA tool described in the first part. However, the data to do so was collected as part of the crossover experiment in the summer of 2020: similar to the design to compare the inter-rater reliability between checkbox grading and traditional grading, the 45 mathematics teachers from the first study also had to regrade all the 60 linear equation tasks from the student under the same SA/PP condition on a self-chosen moment from home

with at least three months between the experiment and the regrading. As such, the intra-rater reliability (degree of agreement among repeated assessments by the same teacher; Parkes, 2012) of SA marking can be compared to PP marking. As teachers connected partial scores to feedback items and could reuse the same items in the regrading process for the SA condition, we hypothesised the intra-reliability of SA scores would be higher than PP scores, as they just had to ‘click together’ the same items as during the experiment.

Nevertheless, this data still needs to be analysed as it proved challenging to link the SA tool with a marking system that returns the total score the teacher intended. Most marking processes are additive in nature: when a student’s solution to a task is assessed, the additive approach initially assumes the student has earned 0 points, and then points are awarded for adherence to the expectations. In contrast, subtractive marking initially assumes students have earned full marks, and then points are subtracted with every mistake, error or omission. The additive approach awards points for how correct an answer is, while the subtractive approach measures how incorrect an answer is. A case study by Becker and Casey (2009) found no significant differences between additive and subtractive marking.

However, teachers often use (unwittingly) a mixture of additive and subtractive approaches. For example, they add points for several correct steps in solving a math problem but subtract a point for a slight miscalculation. In the presented SA system of the first part, feedback items can be associated with points to be added, subtracted or with a maximum. When teachers blend the additive and subtractive approach, the *adder/subtractor problem* arises. Indeed, when feedback in the SA system consists of feedback items adding points, items associated with minus points, and items holding a threshold, the system can not unambiguously decide whether to start from 0, the full mark or the threshold.

Atomic feedback	Full mark: 10 points
• Big misconception	
	Threshold: Max 5/10
• Well executed step	+2 points
• Well executed step	+1 points
• Small mistake	-1 points
	Grade: 4/10, 2/10 or 7/10?

Figure 8 – Which total score is appropriate? When addition, subtraction and thresholds are mixed, the *adder/subtractor problem* is triggered.

For example, in [Figure 8](#), the final mark has to be calculated based on a threshold leading to a maximum of 5/10, two additions (+2 and +1) and one subtraction (-1); this composition can lead to a final mark of 5/10 ($5-1+1+2 = 7$, but the threshold can not be exceeded), of 2/10 ($2+1-1$, 2/10 does not exceed the maximum threshold so no need to take it explicitly into account), of 4/10 (threshold - 1, the two additions are omitted due to presence of a maximum threshold), or 7/10 (the additions and subtractions add up to +2, which is added as a bonus to the threshold).

The adder/subtractor problem is, in nature, a non-deterministic problem. Therefore, it is not straightforward to solve it. However, the SA system would be much more consistent with teachers' grading habits if it allowed this blend of addition and subtraction. During the crossover experiment, a naive algorithm calculated the total score for a question based on the partial scores linked to the feedback items. Teachers could overrule this proposed total score. Analysing these overhauls is an exciting idea for further research. It will give insights into mathematics teachers' additive or subtractive grading approaches, but it is also a required first step to link the SA tool from part 1 to an intuitive marking system. Only after these steps are taken the question of intra-rater reliability can be adequately answered. It would otherwise compare a sometimes unintuitive marking system (SA) with teachers' intuitive approaches to marking they use unconsciously every day (PP).

Combining semi-automated assessment with Bayesian networks

Besides investigating links with grading systems, semi-automated assessment approaches open possibilities to extensively monitor students' individual learning process (i.e. student tracking) and use this information to apply adapted differentiated instruction (cf. adaptive sequencing).

In order to process student tracking data, a Bayesian network scoring engine (Almond et al., 2015) can be used. Bayesian networks are graphical probability models linking latent proficiency variables to the outcomes of a task: each atomic feedback item a teacher selects provides information to update the Bayesian network. All the information collected in the Bayesian network can eventually provide a predictive probability of a student's proficiency in a particular mathematical topic. This information can then be used to guide instruction. For instance, when the Bayesian network indicates a low probability of solving word problems using derivatives, the teacher can purposefully support the student by offering extra information, exercises etc. This is also called adapted differentiated instruction and the Bayesian network allows us to determine which competencies the teacher needs more evidence of the student's mastery and decide which next problem to offer a student.

For example, during the ACED-project (Shute et al., 2007), an intelligent tutoring system with a Bayesian scoring engine was designed on the mathematical topic of 'algebraic sequences'. The system provided informative feedback and adaptive task selection based on a Bayesian network, significantly improving the student's ability.

The idea of correcting handwritten tasks using a semi-automated monitoring system based on Bayesian networks has not yet been studied. We hypothesise that a SA monitoring system that enables student tracking and adaptive sequencing will lead to more effective student learning than simple paper-based assessments.

Comparing atomic with classic feedback in a formative classroom setting

The graphical outline of this dissertation in [Figure 6](#) reveals an important research gap: for the first semi-automated approach, SA and PP feedback was compared using text mining and a qualitative analysis giving some impression on the given feedback, but feedback quality is tested best in the eye of the beholder: the students. A potential study design to do so is by using two-stage tasks. First, students make a challenging mathematics task. Next, the teacher gives feedback to it in two different ways: half of the students get atomic, reusable feedback; the other half get classic written feedback. The students are randomised across these conditions. Next, students should improve their tasks based on the given feedback. These tasks should be graded in both stages to see if one feedback type outperforms the other. In addition, some students could be invited to execute their task improvement during an interview using a think-aloud protocol. While it is a nice research design, there are many confounding factors to consider, making it quite challenging to turn it into a scientifically sound experiment. Moreover, a null result is likely: it may well be that the feedback style does not matter so much as long as students are given some pointers on what they need to improve.

Using checkbox grading for peer feedback and self-assessment

To enhance students' assessment literacy (Winstone et al., 2017), an interesting idea was given to us during a conference: the checkbox grading approach could be suited for peer assessment in mathematics classrooms. In such a study, students could assess each other (or themselves!) by filling out checkbox grading schemes. This could give students an in-depth understanding of which aspects are considered important in challenging mathematics tasks.

BIBLIOGRAPHY

- 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>
- Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). The future of bayesian networks in educational assessment. In *Bayesian networks in educational assessment* (pp. 583–599). Springer New York.
- American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders*. <https://doi.org/10.1176/appi.books.9780890425787>
- Anni, C. T., Sunawan, & Haryono. (2018). School counselors' intention to use technology: The technology acceptance model. *Turkish Online Journal of Educational Technology*, 17, 120–124.
- Artelt, C., & Rausch, T. (2014). Accuracy of teacher judgments. In S. Krolak-Schwerdt, S. Glock, & M. Böhmer (Eds.), *Teachers' professional development: Assessment, training, and learning* (pp. 27–43). SensePublishers. https://doi.org/10.1007/978-94-6209-536-6_3
- 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*, 29(5). <https://doi.org/10.14742/ajet.526>
- Averell, L., & Heathcote, A. (2011). The form of the forgetting curve and the fate of memories [Special Issue on Hierarchical Bayesian Models]. *Journal of Mathematical Psychology*, 55(1), 25–35. <https://doi.org/https://doi.org/10.1016/j.jmp.2010.08.009>
- Backes, B., & Cowan, J. (2019). Is the pen mightier than the keyboard? the effect of online testing on measured student achievement. *Economics of Education Review*, 68, 89–103. <https://doi.org/https://doi.org/10.1016/j.econedurev.2018.12.007>
- Baird, J.-A., 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>
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychol. Rev.*, 84(2), 191–215. [https://doi.org/https://doi.org/10.1016/0146-6402\(78\)90002-4](https://doi.org/https://doi.org/10.1016/0146-6402(78)90002-4)
- Bazvand, A. D., & 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>
- Becker, B. A., & Casey, K. (2009). Half empty, half full - an examination of subtractive versus additive assessment. *Irish Conference on Engaging Pedagogy (ICEP 2009)*. <https://mural.maynoothuniversity.ie/10181/>

- Benett, R. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5–25. <https://doi.org/10.1080/0969594X.2010.513678>
- Bennett, E., Alpert, R., & Goldstein, A. (1954). Communications Through Limited-Response Questioning*. *Public Opinion Quarterly*, 18(3), 303, 308. <https://doi.org/10.1086/266520>
- Benoit, K., Muhr, D., & Watanabe, K. (2021). *Stopwords: Multilingual stopword lists*. <https://CRAN.R-project.org/package=stopwords>
- 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. v. d. Heuvel-Panhuizen, & M. Veldhuis (Eds.), *Proceedings of the eleventh congress of the european society for research in mathematics education (CERME11)* (Vol. TWG21). Freudenthal Group. <https://hal.science/hal-02430515>
- Bose, M., & Dey, A. (2009). *Optimal crossover designs*. World Scientific. <https://doi.org/10.1142/6878>
- Brooks, V. (2004). Double marking revisited. *British Journal of Educational Studies*, 52(1), 29–46. <https://doi.org/10.1111/j.1467-8527.2004.00253.x>
- Brunner, M., & Süß, 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>
- Bugbee, A. C., Jr. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, 28(3), 282–299.
- Burkhardt, H. (1985). Curricula for active mathematics: Developments in school mathematics education around the world. In I. Wirszup & R. Streit (Eds.), *Proceedings of the ucsm international conference on mathematics education* (pp. 321–361).
- Busch, J., Barzel, B., & Leuders, T. (2015a). Promoting secondary teachers' diagnostic competence with respect to functions: Development of a scalable unit in continuous professional development. *ZDM - Mathematics Education*, 47(1), 53–64. <https://doi.org/10.1007/s11858-014-0647-2>
- Busch, J., Barzel, B., & Leuders, T. (2015b). Die Entwicklung eines Instruments zur kategorialen Beurteilung der Entwicklung diagnostischer Kompetenzen von Lehrkräften im Bereich Funktionen. *Journal für Mathematik-Didaktik*, 36(2), 315–338. <https://doi.org/10.1007/s13138-015-0079-8>
- Candel, C., Vidal-Abarca, E., Cerdán, R., Lippmann, M., & Narciss, S. (2020). Effects of timing of formative feedback in computer-assisted learning environments. *Journal of Computer Assisted Learning*, 36(5), 718–728. <https://doi.org/10.1111/jcal.12439>
- 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>
- Cassidy, S., & Eachus, P. (2002). Developing the computer user self-efficacy (cuse) scale: Investigating the relationship between computer self-efficacy, gender and experience with computers. *Journal of Educational Computing Research*, 26(2), 133–153. <https://doi.org/10.2190/jgjr-0kvl-hrf7-gcnv>
- Chang, N., Watson, A. B., Bakerson, M. A., Williams, E. E., McGoron, F. X., & Spitzer, B. (2012). Electronic feedback or handwritten feedback: What do undergraduate students prefer and why? [Section: Articles]. *Journal of Teaching and Learning with Technology*, 1(1), 1–23. Retrieved January 30, 2023, from <https://scholarworks.iu.edu/journals/index.php/jotlt/article/view/2043>
- Chiles, M. (2021). *The feedback pendulum* [OCLC: on1163957811]. John Catt Educational Ltd.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge. <https://doi.org/10.4324/9780203771587>
- Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88(2), 322–328. <https://doi.org/10.1037/0033-2909.88.2.322>
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology [Publisher: Management Information Systems Research Center, University of Minnesota]. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>
- Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8), 982–1003. <https://doi.org/10.1287/mnsc.35.8.982>
- 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>
- De Smedt, T., & Daelemans, W. (2012). “vreselijk mooi!” (terribly beautiful): A subjectivity lexicon for dutch adjectives. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, 3568–3572. http://www.lrec-conf.org/proceedings/lrec2012/pdf/312_Paper.pdf
- De Vries, H., Elliott, M. N., Kanouse, D. E., & Teleki, S. S. (2008). Using pooled kappa to summarize interrater agreement across many items. *Field Methods*, 20(3), 272–282. <https://doi.org/10.1177/1525822x08317166>
- Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4), 227–268. https://doi.org/https://doi.org/10.1207/S15327965PLI1104_01
- Denton, P., Madden, J., Roberts, M., & Rowe, P. (2008). Students’ response to traditional and computer-assisted formative feedback: A comparative case study. *British Journal of Educational Technology*, 39(3), 486–500. <https://doi.org/https://doi.org/10.1111/j.1467-8535.2007.00745.x>
- Denton, P., & McIlroy, D. (2018). Response of students to statement bank feedback: The impact of assessment literacy on performances in summative tasks. *Assessment & Evaluation in Higher Education*, 43(2), 197–206. <https://doi.org/10.1080/02602938.2017.1324017>
- Denton, P., & Rowe, P. (2015). Using statement banks to return online feedback: Limitations of the transmission approach in a credit-bearing assessment. *Assessment & Evaluation in Higher Education*, 40(8), 1095–1103. <https://doi.org/10.1080/02602938.2014.970124>

- Depaepe, F., Verschaffel, L., & Kelchtermans, G. (2013). Pedagogical content knowledge: A systematic review of the way in which the concept has pervaded mathematics educational research. *Teaching and Teacher Education*, 34, 12–25. <https://doi.org/10.1016/j.tate.2013.03.001>
- Docktor, J., & Heller, K. (2009). Robust assessment instrument for student problem solving. *Proceedings of the NARST 2009 Annual Meeting*, Garden Grove, CA, 1(1), 1–19.
- Drijvers, P. (2018). *Hoofd in de wolken, voeten op de vloer. Praktijkgericht onderzoek naar wiskundig denken in ICT-rijk wiskundeonderwijs. (Dutch) [Head in the clouds, feet on the floor. Practice-based research on mathematical thinking in ICT-rich mathematics education.]* Kenniscentrum leren en innoveren, Hogeschool Utrecht.
- Enu, J. (2021). Factors affecting teacher educators adoption of formative assessment strategies in the mathematics classroom. *Journal of Education and Learning (EduLearn)*, 15(4), 483–489. <https://doi.org/10.11591/edulearn.v15i4.20341>
- Eurydice. (2021). *Teachers in europe: Careers, development and well being. eurydice report.* Luxembourg: Publications Office of the European Union. <https://doi.org/10.2797/997402>
- Eva, K. W., Armson, H., Holmboe, E., Lockyer, J., Loney, E., Mann, K., & Sargeant, J. (2011). Factors influencing responsiveness to feedback: On the interplay between fear, confidence, and reasoning processes. *Advances in Health Sciences Education*, 17(1), 15–26. <https://doi.org/10.1007/s10459-011-9290-7>
- Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research*, 83(1), 70–120. <https://doi.org/10.3102/0034654312474350>
- Evans, J. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59(1), 255–278. <https://doi.org/10.1146/annurev.psych.59.103006.093629>
- Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. *Dental Press Journal of Orthodontics*, 19(4), 27–29. <https://doi.org/10.1590/2176-9451.19.4.027-029.ebo>
- Fahlgren, M., Brunström, M., Dilling, F., Kristinsdóttir, B., Pinkernell, G., & Weigand, H.-G. (2021). Technology-rich assessment in mathematics. In *Mathematics education in the digital age* (pp. 69–83). Routledge. <https://doi.org/10.4324/9781003137580-5>
- Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6), 543–549. [https://doi.org/10.1016/0895-4356\(90\)90158-L](https://doi.org/10.1016/0895-4356(90)90158-L)
- Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). Text mining in education. *WIREs Data Mining and Knowledge Discovery*, 9(6). <https://doi.org/10.1002/widm.1332>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5), 378. <https://doi.org/10.1037/h0031619>
- Fleiss, J. L., Nee, J. C., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86(5), 974–977. <https://doi.org/10.1037/0033-2909.86.5.974>
- Flemish education inspectorate. (2023). *Onderwijsspiegel 2023. Jaarlijks rapport van de onderwijsinspectie. (Dutch) [Year report 2023 of the Flemish education inspectorate.]* Vlaamse Overheid. <https://www.onderwijsinspectie.be/nl/andere-opdrachten/andere/jaarverslag-onderwijsspiegel>
- 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. (2009). *The checklist manifesto.* Macmillan USA.

- Gibbs, G., & Simpson, C. (2005). Conditions under which assessment supports students' learning [Number: 1 Publisher: University of Gloucestershire]. *Learning and Teaching in Higher Education*, (1), 3–31. Retrieved January 30, 2023, from <https://eprints.glos.ac.uk/3609/>
- Gibson, S., Oliver, L., & Dennison, M. (2015). *Workload challenge: Analysis of teacher consultation responses* (Vol. 355). Department for Education London.
- Gillham, B. (2005). *Research interviewing : The range of techniques*. Open University Press.
- Gleaves, A., & Walker, C. (2013). Richness, redundancy or relational salience? a comparison of the effect of textual and aural feedback modes on knowledge elaboration in higher education students' work. *Computers & Education*, 62, 249–261. <https://doi.org/10.1016/j.compedu.2012.11.004>
- Glover, C., & Brown, E. (2006). Written feedback for students: Too much, too detailed or too incomprehensible to be effective? *Bioscience Education*, 7(1), 1–16. <https://doi.org/10.3108/beej.2006.07000004>
- Goetz, T., Lipnevich, A. A., Krannich, M., & Gogol, K. (2018). Performance feedback and emotions. In *The cambridge handbook of instructional feedback* (pp. 554–574). Cambridge University Press. <https://doi.org/10.1017/9781316832134.027>
- Google. (2021). *Give feedback on assignments in google classroom*. <https://support.google.com/edu/classroom/answer/9093530?hl=en#zippy=%2Cadd-and-save-comments-to-use-later>.
- Gravemeijer, K., Stephan, M., Julie, C., Lin, F.-L., & Ohtani, M. (2017). What mathematics education may prepare students for the society of the future? *International Journal of Science and Mathematics Education*, 15(S1), 105–123. <https://doi.org/10.1007/s10763-017-9814-6>
- Griffin, P., & Care, E. (Eds.). (2015). *Assessment and teaching of 21st century skills*. Springer Netherlands. <https://doi.org/10.1007/978-94-017-9395-7>
- Grönberg, N., Knutas, A., Hynninen, T., & Hujala, M. (2021). Palaute: An online text mining tool for analyzing written student course feedback. *IEEE Access*, 9, 134518–134529. <https://doi.org/10.1109/access.2021.3116425>
- Gusukuma, L., Bart, A. C., Kafura, D., & Ernst, J. (2018). Misconception-driven feedback: Results from an experimental study. *Proceedings of the 2018 ACM Conference on International Computing Education Research*, 160–168. <https://doi.org/10.1145/3230977.3231002>
- Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *The British journal of mathematical and statistical psychology*, 61, 29–48. <https://doi.org/10.1348/000711006X126600>
- Gwet, K. L. (2012). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.
- Gwet, K. L. (2021). Large-sample variance of fleiss generalized kappa. *Educational and Psychological Measurement*, 81(4), 781–790. <https://doi.org/10.1177/0013164420973080>
- Hanna, R., & Linden, L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy*, 4(4), 146–168. <https://doi.org/10.1257/pol.4.4.146>
- 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>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77, 81–112. <https://doi.org/10.3102/003465430298487>
- 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>

- Higgins, R., Hartley, P., & Skelton, A. (2001). Getting the message across: The problem of communicating assessment feedback. *Teaching in Higher Education*, 6(2), 269–274. <https://doi.org/https://doi.org/10.1080/13562510120045230>
- 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>
- Huang, S. (2000). A quantitative analysis of audiotaped and written feedback produced for students' writing and students' perceptions of the two feedback methods. *Tunghai Journal*, 41, 199–232.
- Hujala, M., Knutas, A., Hynninen, T., & Arminen, H. (2020). Improving the quality of teaching by utilising written student feedback: A streamlined process. *Computers & Education*, 157, 103965. <https://doi.org/https://doi.org/10.1016/j.compedu.2020.103965>
- Hull, M. M., Kuo, E., Gupta, A., & Elby, A. (2013). Problem-solving rubrics revisited: Attending to the blending of informal conceptual and formal mathematical reasoning. *Physical Review Special Topics - Physics Education Research*, 9(1). <https://doi.org/10.1103/physrevstper.9.010105>
- Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, 41(2), 201–213. <https://doi.org/https://doi.org/10.2307/358160>
- 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>
- Junqueira, L., & Payant, C. (2015). “i just want to do it right, but it's so hard”: A novice teacher's written feedback beliefs and practices. *Journal of Second Language Writing*, 27, 19–36. <https://doi.org/https://doi.org/10.1016/j.jslw.2014.11.001>
- Keller, J. M. (2009). Motivational design research and development. In *Motivational design for learning and performance* (pp. 297–323). Springer US. https://doi.org/10.1007/978-1-4419-1250-3_12
- Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up*. Washington, DC: National Research Council. <https://doi.org/10.17226/9822>
- Kinnear, G., Jones, I., Sangwin, C., Alarfaj, M., Davies, B., Fearn, S., Foster, C., Heck, A., Henderson, K., Hunt, T., Iannone, P., Kontorovich, I., Larson, N., Lowe, T., Meyer, J. C., O'Shea, A., Rowlett, P., Sikurajapathi, I., & Wong, T. (2022). A collaboratively-derived research agenda for e-assessment in undergraduate mathematics. *International Journal of Research in Undergraduate Mathematics Education*. <https://doi.org/10.1007/s40753-022-00189-6>
- Kirschner, P. A., van den Brink, H., & Meester, M. (1991). Audiotape feedback for essays in distance education. *Innovative Higher Education*, 15(2), 185–195. <https://doi.org/10.1007/bf00898030>
- 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>

- Kraemer, H. C. (1980). Extension of the kappa coefficient. *Biometrics*, 36(2), 207–216. <https://doi.org/10.2307/2529972>
- Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine*, 21(14), 2109–2129. <https://doi.org/10.1002/sim.1180>
- Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research*, 30(3), 411–433. <https://doi.org/10.1111/j.1468-2958.2004.tb00738.x>
- Kuhlemeier, H., van Rijn, P., & Kremers, E. (2012). *Eerste, tweede en derde correctie van examenwerk: Wat is het verschil? (Dutch) [First, second and third assessment of an exam: What is the difference?]* Cito, Arnhem. <https://www.cito.nl/-/media/files/kennisbank/psychometrie/artikelen-in-tijdschriften/cit0a29kuhlemeier-van-rijn-kremers-2013-rapportage-eerste-tweede-en-derde-correctie.pdf>
- Kwartler, T. (2017). *Text mining in practice with r*. John Wiley & Sons.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159. <https://doi.org/10.2307/2529310>
- Lefevre, D., & Cox, B. (2017). Delayed instructional feedback may be more effective, but is this contrary to learners' preferences?: Timing of TBI feedback. *British Journal of Educational Technology*, 48(6), 1357–1367. <https://doi.org/10.1111/bjet.12495>
- 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>
- Lim, K. H., Benbasat, I., & Todd, P. A. (1996). An experimental investigation of the interactive effects of interface style, instructions, and task familiarity on user performance. *ACM Transactions on Computer-Human Interaction*, 3(1), 1–37. <https://doi.org/10.1145/226159.226160>
- Lipnevich, A. A., Berg, D. A., & Smith, J. K. (2016). Toward a model of student response to feedback. In *Handbook of human and social conditions in assessment* (pp. 169–185). 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>
- Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404), 1198–1202. <https://doi.org/10.1080/01621459.1988.10478722>
- Locke, E. A., & Latham, G. P. (2013). Goal setting theory, 1990. In *New developments in goal setting and task performance*. Routledge. <https://doi.org/10.4324/9780203082744>
- Loibl, K., Leuders, T., & Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgments by cognitive modeling (DiaCoM). *Teaching and Teacher Education*, 91, 103059. <https://doi.org/10.1016/j.tate.2020.103059>
- MacLure, M. (2013). Chapter 9 classification or wonder? coding as an analytic practice in qualitative research. In *Deleuze and research methodologies* (pp. 164–183). Edinburgh University Press. <https://doi.org/10.1515/9780748644124-011>
- McMillan, J. (2013). *SAGE handbook of research on classroom assessment*. SAGE Publications, Inc. <https://doi.org/10.4135/9781452218649>
- Meadows, M., & Billington, L. (2005). A review of the literature on marking reliability. https://filestore.aqa.org.uk/content/research/CERP_RP_MM_01052005.pdf
- Mezzich, J. E., Kraemer, H. C., Worthington, D. R., & Coffman, G. A. (1981). Assessment of agreement among several raters formulating multiple diagnoses. *Journal of Psychiatric Research*, 16(1), 29–39. [https://doi.org/10.1016/0022-3956\(81\)90011-X](https://doi.org/10.1016/0022-3956(81)90011-X)

- Mohanty, S., Chatterjee, J., Jain, S., Elngar, A., & Gupta, P. (2020). *Recommender system with machine learning and artificial intelligence: Practical tools and applications in medical, agricultural and other industries*. Wiley. <https://doi.org/10.1002/9781119711582>
- Moodle. (2022). *Advanced grading methods* (tech. rep.). Moodle. https://docs.moodle.org/402/en/Advanced_grading_methods
- Moodle. (2023). *Moodle developer documentation* (tech. rep.). Moodle. https://docs.moodle.org/dev/Main_Page
- Moons, F., & Vandervieren, E. (2022). Handwritten math exams with multiple assessors: researching the added value of semi-automated assessment with atomic feedback. *Twelfth Congress of the European Society for Research in Mathematics Education (CERME12), TWG21(14)*. <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. <https://doi.org/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 and Education Open*, 3, 100086. <https://doi.org/10.1016/j.caeo.2022.100086>
- Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research, and Evaluation*, 7. <https://doi.org/10.7275/A5VQ-7Q66>
- Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research, and Evaluation*, 7(10). <https://doi.org/https://doi.org/10.7275/q7rm-gg74>
- Movshovitz-Hadar, N., Zaslavsky, O., & Inbar, S. (1987). An empirical classification model for errors in high school mathematics [Publisher: National Council of Teachers of Mathematics]. *Journal for Research in Mathematics Education*, 18(1), 3–14. <https://doi.org/10.2307/749532>
- NVivo. (2022). *Run a coding comparison query* (tech. rep.). NVivo 11. https://help-nv11.qsrinternational.com/desktop/procedures/run_a_coding_comparison_query.htm
- 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>
- OECD. (2012). How many students are in each classroom? [Series Title: Highlights from Education at a Glance]. In *Education at a glance 2012*. https://doi.org/10.1787/eag_highlights-2012-25-en
- 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>
- Osgood, C. E. (1959). The representational model and relevant research methods. In I. Pool (Ed.), *Trends in content analysis* (pp. 33, 38). Urbana: University of Illinois Press.
- Panadero, E., Lipnevich, A., & Broadbent, J. (2019). Turning self-assessment into self-feedback. In *The impact of feedback in higher education* (pp. 147–163). Springer International Publishing. https://doi.org/10.1007/978-3-030-25112-3_9
- Parkes, J. (2012). *Reliability in classroom assessment* (J. H. McMillan, Ed.) [Pages: 124 Publication Title: SAGE Handbook of Research on Classroom Assessment]. SAGE Publications.

- Pearce, C. G., & Ackley, R. J. (1995). Audiotaped feedback in business writing: An exploratory study. *Business Communication Quarterly*, 58(3), 31–34. <https://doi.org/10.1177/108056999505800306>
- Pelkola, T., Rasila, A., & Sangwin, C. (2018). Investigating bloom's learning for mastery in mathematics with online assessment. *Informatics in Education*, 17(2), 363–380. <https://doi.org/10.15388/infedu.2018.19>
- Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback : All that effort, but what is the effect? *Assessment & Evaluation in Higher Education*, 35(3), 277–289. <https://doi.org/10.1080/02602930903541007>
- Price, M., Rust, C., O'Donovan, B., Handley, K., & Bryant, R. (2012). *Assessment literacy: The foundation for improving student learning*. Oxford Centre for Staff; Learning Development.
- Rakoczy, K., Harks, B., Klieme, E., Blum, W., & Hochweber, J. (2013). Written feedback in mathematics: Mediated by students' perception, moderated by goal orientation. *Learning and Instruction*, 27, 63–73. <https://doi.org/10.1016/j.learninstruc.2013.03.002>
- Ratkowsky, D. A., Evans, M. A., & Alldredge, J. R. (1993). *Cross-over experiments: Design, analysis and application*. Dekker.
- Reusser, K., & Stebler, R. (1997). Every word problem has a solution—the social rationality of mathematical modeling in schools. *Learning and Instruction*, 7(4), 309–327. [https://doi.org/10.1016/S0959-4752\(97\)00014-5](https://doi.org/10.1016/S0959-4752(97)00014-5)
- Ripley, M. (2009). Transformational computer-based testing. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment. new approaches to skills assessment and implications for large-scale testing*. (pp. 92–98). Luxembourg; Office for Official Publications of the European Communities. <https://doi.org/https://doi.org/10.2788/60083>
- 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>
- Ryan, T., Henderson, M., & Phillips, M. (2019). Feedback modes matter: Comparing student perceptions of digital and non-digital feedback modes in higher education. *British Journal of Educational Technology*, 50(3), 1507–1523. <https://doi.org/10.1111/bjet.12749>
- Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, 34(2), 159–179. <https://doi.org/https://doi.org/10.1080/02602930801956059>
- Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, 35(5), 535–550. <https://doi.org/10.1080/02602930903541015>
- Sangwin, C. (2013a). The future. In *Computer aided assessment of mathematics*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199660353.003.0009>
- Sangwin, C. (2013b). *Computer aided assessment of mathematics*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199660353.001.0001>
- Sangwin, C., & Jones, I. (2017). Asymmetry in student achievement on multiple-choice and constructed-response items in reversible mathematics processes. *Educational Studies in Mathematics*, 94(2), 205–222. <https://doi.org/https://doi.org/10.1007/s10649-016-9725-4>
- Sangwin, C., & Köcher, N. (2016). Automation of mathematics examinations. *Computers & Education*, 94, 215–227. <https://doi.org/https://doi.org/10.1016/j.compedu.2015.11.014>
- Scherer, R., Siddiq, F., & Tondeur, J. (2019). The technology acceptance model (TAM): A meta-analytic structural equation modeling approach to explaining teachers' adoption of

- digital technology in education. *Computers & Education*, 128, 13–35. <https://doi.org/https://doi.org/10.1016/j.compedu.2018.09.009>
- Schnepper, L. C., & McCoy, L. P. (2014). Analysis of misconceptions in high school mathematics. *Networks: An Online Journal for Teacher Research*, 15(1). <https://doi.org/https://doi.org/10.4148/2470-6353.1066>
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189. <https://doi.org/10.3102/0034654307313795>
- Shute, V. J., Hansen, E. G., & Almond, R. G. (2007). An assessment for learning system called aced: Designing for learning effectiveness and accessibility. *ETS Research Report Series*, 2007(2), i–45. <https://doi.org/10.1002/j.2333-8504.2007.tb02068.x>
- Silge, J., & Robinson, D. (2017). *Text mining with r: A tidy approach*. O'Reilly.
- Singh, A., Karayev, S., Gutowski, K., & Abbeel, P. (2017). Gradescope. *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*. <https://doi.org/10.1145/3051457.3051466>
- Stanyon, R., Martello, E., Kainth, M., & Wilkin, N. K. (2022). Demo of graide: AI powered assistive grading engine. *Proceedings of the Ninth ACM Conference on Learning @ Scale*. <https://doi.org/10.1145/3491140.3528263>
- Starch, D., & Elliott, E. C. (1913). Reliability of grading work in mathematics. *The School Review*, 21(4), 254–259. Retrieved April 29, 2023, from <http://www.jstor.org/stable/1076246>
- Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey, A. R., & Schmitz, J. A. P. (2009). An assessment of reliability and validity of a rubric for grading APA-style introductions. *Teaching of Psychology*, 36(2), 102–107. <https://doi.org/10.1080/00986280902739776>
- Taber, K. S. (2018). The use of cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(6), 1273–1296. <https://doi.org/https://doi.org/10.1007/s11165-016-9602-2>
- 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>
- Thurm, D., & Barzel, B. (2021). Teaching mathematics with technology: A multidimensional analysis of teacher beliefs. *Educational Studies in Mathematics*, 109(1), 41–63. <https://doi.org/10.1007/s10649-021-10072-x>
- Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis [Publisher: American Educational Research Association]. *Review of Educational Research*, 85(4), 475–511. <https://doi.org/10.3102/0034654314564881>
- van de Walle, J., Karp, K., & Bay-Williams, J. (2018). *Elementary and middle school mathematics: Teaching developmentally*. Pearson.
- Vanacore, A., & Pellegrino, M. S. (2022). Benchmarking procedures for characterizing the extent of rater agreement: A comparative study. *Quality and Reliability Engineering International*, 38(3), 1404–1415. <https://doi.org/10.1002/qre.2982>
- VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16(3), 227–265.
- Venkatesh, Morris, Davis, & Davis. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425. <https://doi.org/10.2307/30036540>
- Warrens, M. J. (2010). A formal proof of a paradox associated with cohen's kappa. *Journal of Classification*, 27(3), 322–332. <https://doi.org/10.1007/s00357-010-9060-x>

-
- 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>
- Weiner, B. (1986). *An attributional theory of motivation and emotion*. Springer US. <https://doi.org/10.1007/978-1-4612-4948-1>
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25(1), 68–81. <https://doi.org/https://doi.org/10.1006/ceps.1999.1015>
- 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>
- Wisniewski, B., Zierer, K., & Hattie, J. (2019). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 3087. <https://doi.org/https://doi.org/10.3389/fpsyg.2019.03087>
- Yan, T., Conrad, F. G., Tourangeau, R., & Couper, M. P. (2010). 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>
- Yang, K.-H., & Lu, B.-C. (2021). Towards the successful game-based learning: Detection and feedback to misconceptions is the key. *Computers & Education*, 160, 104033. <https://doi.org/10.1016/j.compedu.2020.104033>
- Yu, C., Jannasch-Pennell, A., & DiGangi, S. (2011). Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. *The Qualitative Report*. <https://doi.org/10.46743/2160-3715/2011.1085>
- Zumbach, D., & Bauer, P. C. (2021). *Deeplr: Interface to the 'DeepL' translation API*. <https://CRAN.R-project.org/package=deeplr>

SUMMARY (ENGLISH)

Feedback is the most powerful engine of any learning process. In the field of mathematics education, the possibilities to assess automatically are therefore being thoroughly explored. However, students face difficulties expressing themselves mathematically on a computer and learning systems can often only assess the outcome and not the solving method. Research indicates that automated tests focus too much on procedural fluency at the expense of higher-order thinking questions. It takes much effort to develop digital tests, and teachers are sceptical of using automated assessments, meaning that paper-and-pencil tests still dominate mathematics classrooms. One of the characteristics of mathematical assessment is that wrong answers tend to exhibit patterns among the student population. Consequently, teachers often repeat their feedback and grades. This brings us to the idea of semi-automated feedback and assessment: by correcting handwritten tasks digitally, feedback can be saved and reused. This could lead to more elaborate feedback, time savings, and enhanced inter-rater reliability. Specifically, two semi-automatic assessment approaches were developed and investigated.

In the first study, a software tool was programmed, allowing teachers to write feedback for a student, and the computer saves it so that it can be reused when subsequent students make the same or similar mistakes. In order to teach educators how to compose reusable feedback, the concept of atomic feedback has been introduced. To write atomic feedback, teachers have to identify the independent errors and write brief feedback items for each separate error. Feedback items that belong together (either thematically or in the solving method) can be clustered to form a hierarchical list of feedback items. It has been shown that these format requirements for mathematical feedback items significantly increase the reusability of feedback. Moreover, teachers quickly adopted them: during a crossover experiment with 45 mathematics teachers, 74% of the given feedback using the semi-automated approach could be categorised as atomic after only a brief introduction to the concept of atomic feedback. The 26% non-atomic items often addressed multiple issues, which must be in separate feedback items in order to be atomic. Including the location in a student's solution where the error occurred was the most common violation. In the spirit of atomic feedback, the location and error should be spread across two feedback items: one for the location and one for the error. They can be glued together in the list of feedback items by indenting the item about the error below the item with the location.

A remarkable result was discovered during the experiment: the semi-automated tool led teachers to give significantly more feedback instead of saving time compared to

classic pen-and-paper feedback. These two feedback types were also compared by form and content using text mining and qualitative techniques. Word frequencies, sentiments and the amount of erroneous, descriptive and corrective feedback were similar in both feedback types. When teachers used the semi-automated tool, the feedback was more elaborate but less specific to the student's solution. Without the tool, feedback was shorter but more concrete and focused on the main issues. The teachers' feedback with the semi-automatic tool did not always have better properties than classic pen-and-paper feedback: teachers tend to describe and correct students' work instead of analysing underlying (mis-)conceptions using it. Overall, teachers may not confuse handiness with quality: having high feedback literacy and pedagogical content knowledge remain essential teachers' characteristics to ensure high-quality feedback in both types.

The second study was conducted in collaboration with the Flemish Exam Commission. Their traditional grading method of handwritten mathematics exams was transformed into a semi-automated one called 'checkbox grading.' Every assessor receives a list of checkboxes, and they must tick those that apply to the student's solution. Dependencies between these checkboxes can be set to ensure all assessors take the same path down the grading scheme. The system automatically calculates the grade and results in atomic feedback giving a detailed insight into what went wrong and how the grade was obtained. In the traditional grading method, the assessors only communicate a grade based on predefined grading schemes provided by the commission. The method was investigated from both assessors' and students' points of view.

From the assessors' perspective, we investigated their time investment, views and inter-rater reliability. Concerning their time investments, checkbox grading took about twice as long as assessing an exam the traditional way. Surprisingly, assessors' subjective feelings of time were at odds with these measurements: they largely agreed that checkbox grading allowed them to perform their duties as an assessor more quickly. It might be that assessors considered the transparency of checkbox grading and the resulting students' feedback more critical than their time commitment. This is supported by their positive views on checkbox grading, in which they reported high perceived usefulness and a strong attitude towards using the semi-automated approach.

Assessors' inter-rater reliability was compared between blind versus visible checkbox grading and against the traditional grading method. In checkbox grading, the checkboxes can be linked to partial scores. By clicking the checkboxes that apply to a student's solution, the computer automatically calculates the grade. This led to the experimental condition of blind checkbox grading, where the underlying grades are not shown to the assessors. From the literature on rubrics, it is known that assessors start deviating from the criteria when their holistic evaluation of the student's work does not align with the evaluation resulting from the rubric. By not showing the grades, we wanted to investigate if we could avoid this cognitive conflict which should result in higher inter-rater reliability. However, there appeared to be no appropriate inter-rater reliability measure to answer this research question. Indeed, no chance-corrected κ coefficient existed that could calculate the reliability of multiple raters, classifying each subject into one-or-more categories. Well-known measures such as Cohen's kappa and Fleiss' kappa require that each subject be assigned exactly one category. Measures dropping this condition were barely known, and the few described did not allow these categories to be hierarchical or differing in importance. This led to the discovery of a generalised

Fleiss' kappa, taking into account all the information checkbox grading provides. The measure allowed us to answer the research question: blind grading enhances inter-rater reliability when grading schemes should be interpreted strictly, while visible checkbox grading is better suited for more complex grading schemes as seeing the scores helps assessors to judge the correctness of their own assessment work. Compared to the traditional method, checkbox grading was as reliable.

The investigate how students perceive the resulting atomic feedback from checkbox grading, a questionnaire was conducted by 36 of the 60 students who took part in the exam on which the research was conducted. Four of them agreed to semi-structured interviews. Students preferred traditional grading over checkbox grading when asked to rank feedback types from more to less comprehensible. However, when interviewed using a think-aloud protocol, students were found to interpret checkbox grading feedback more easily. Moreover, 97% students agreed on the questionnaire that the Flemish Exam Commission should adopt the method. The clarity the feedback offers into how their scores were obtained was regarded highly. The student's understanding was high on average and could not be correlated with their exam score, which means that almost all students — both the high- and low-performing ones — could make sense of checkbox grading feedback.

The two semi-automated approaches showed two valuable ways in which computers and teachers can work together to assess and give feedback to students in mathematics education. The first approach was built for individual teachers, and the second was for a group of assessors. Both studies show that giving feedback always requires more work than just highlighting mistakes or communicating a grade and that such tools often motivate teachers to do even more working instead of saving time. We conclude with some ideas for further research. The first priority is making the suggestion system of the first semi-automated approach more intelligent. By incorporating ideas from the literature around recommendation systems, selecting feedback items to reuse could become a much smoother process. Next, linking semi-automated approaches to Bayesian networks would allow us to monitor students' learning processes more sharply. A Bayesian network is a probabilistic graphical model of a student's proficiency. Exploring semi-automated assessment in other settings than done in this dissertation (e.g., peer feedback, formative settings) are other fruitful ideas for further research.

✓ **SAMENVATTING (NEDERLANDS)**

Feedback is de krachtigste motor van elk leerproces. In de wiskundendidactiek wordt daarom uitgebreid onderzocht hoe men beoordelingen kan automatiseren. Dat is niet evident voor leerlingen: zich wiskundig uitdrukken is moeilijk op een computer en leersystemen kunnen vaak enkel de uitkomst verwerken en niet de oplossingsmethode. Digitale testen blijken zich veelal te beperken tot procedurele kennis ten koste van inzichtelijke denkvragen. Digitale wiskundetests ontwikkelen is een tijdrovende klus en daarnaast zijn leraars erg sceptisch om ze in te zetten, waardoor pen-en-papier het wiskundeonderwijs nog steeds domineert. Eén van de karakteristieken van wiskundig beoordelingswerk is dat foute antwoorden in een klasgroep patronen vertonen. Bijgevolg moeten leerkrachten hun feedback en punten meermaals herhalen. Dat brengt ons op het idee van semi-automatisch beoordelen: door handgeschreven leerlingoplossingen digitaal te beoordelen, kan feedback bewaard en vervolgens hergebruikt worden. Dat kan uitgebreidere feedback, tijdswinst en verhoogde interbeoordelaarsbetrouwbaarheid opleveren. In dit proefschrift werden twee semi-automatische beoordelingsmethoden ontwikkeld en onderzocht.

Voor de eerste studie werd een softwaretool geprogrammeerd die de feedback van de leerkrachten tijdens het nakijken opslaat, zodat die makkelijk hergebruikt kan worden als dezelfde of gelijkaardige fouten zich opnieuw aandienen. Om leerkrachten te leren hoe zij herbruikbare feedback kunnen opstellen, werd atomische feedback uitgevonden. Om feedback atomisch te schrijven, moeten de leraars de verschillende fouten van een student identificeren en korte feedbackitems schrijven voor elke afzonderlijke fout. Hierbij is het van belang dat deze items onafhankelijk van elkaar zijn. Items die thematisch of in de oplossing van de leerling bij elkaar horen, kunnen worden geclusterd tot een hiërarchische lijst. Tijdens het onderzoek konden we aantonen dat atomische feedback de herbruikbaarheid van feedback significant verbetert. Bovendien konden leraars zich deze vormvereisten snel eigen maken: na slechts een korte introductie, kon tijdens een cross-overexperiment met 45 wiskundeleerkrachten, 74% van de gegeven feedback met de feedbacktool als atomisch geclassificeerd worden. De 26% niet-atomische items behandelden vaak meerdere fouten binnen één item of vermeldten zowel de fout als de plaats ervan in de oplossing van de leerling. In de geest van atomische feedback, moeten deze twee verdeeld worden over twee feedbackitems: één voor de locatie en één voor de fout. Ze kunnen in de lijst van feedbackitems aan elkaar worden gelinkt door het item over de fout te laten inspringen onder het item met de locatie.

Tijdens het experiment werd een opvallende ontdekking gedaan: de semi-automatische feedbacktool leidde ertoe dat leerkrachten aanzienlijk meer feedback gaven in plaats van tijd te besparen in vergelijking met het opstellen van klassieke, handgeschreven feedback. Deze twee feedbacktypes werden ook vergeleken naar vorm en inhoud met behulp van text mining en kwalitatieve technieken. Woordfrequenties, gevoelens en de hoeveelheid foutieve, beschrijvende en corrigerende feedback waren vergelijkbaar in beide feedbacktypes. Wanneer leerkrachten de semi-automatische tool gebruikten, was de feedback uitgebreider maar minder specifiek gericht op de oplossing van de leerling. Zonder de tool was de feedback korter, concreter en meer gefocust op de hoofdzaken. De eigenschappen van de feedback met de semi-automatische tool waren zeker niet altijd beter dan die van de klassieke feedback: de leerkrachten hadden vaker de neiging het werk van de leerlingen te beschrijven en te corrigeren in plaats van de onderliggende misconcepties te analyseren. In het algemeen mogen leerkrachten de handigheid van de tool niet verwarren met kwaliteit: een grote feedbackvaardigheid, vakinhoudelijke en vakdidactische kennis blijven essentiële leerkrachtkenmerken om kwaliteitsvolle feedback te garanderen.

De tweede studie was een samenwerking met de Examencommissie Secundair Onderwijs van de Vlaamse overheid. Tijdens het onderzoek werd hun traditionele beoordelingsmethode voor handgeschreven wiskunde-examens omgevormd tot een semi-geautomatiseerd systeem die omgedoopt werd tot 'checkbox grading.' Elke corrector ontvangt een lijst met aankruisvakjes en moet diegene aanvinken die van toepassing zijn op de leerlingenoplossing. Er kunnen afhankelijkheden tussen deze vakjes worden gedefinieerd om ervoor te zorgen dat alle beoordelaars dezelfde weg afleggen doorheen het beoordelingsschema. Het systeem berekent automatisch het cijfer en genereert feedback die een gedetailleerd inzicht geeft in wat er fout ging en hoe het cijfer werd beoaald, gebaseerd op vooraf opgestelde atomische feedback. Bij de traditionele beoordelingsmethode delen de beoordelaars alleen een cijfer mee op basis van beoordelingsschema's die door de examencommissie werden opgesteld. De methode werd onderzocht zowel vanuit het oogpunt van de correctoren, als vanuit het oogpunt van de leerlingen (kandidaten).

We onderzochten de tijdsbesteding, waardering en interbeoordelaarsbetrouwbaarheid van de correctoren. Het nakijken met 'checkbox grading' duurde ongeveer twee keer zo lang als het beoordelen op de traditionele manier. Verrassend genoeg was het subjectieve tijdsbesef van de correctoren in tegenspraak met deze metingen: zij rapporteerden net dat ze sneller hun taken als corrector konden uitvoeren met 'checkbox grading'. Het is mogelijk dat de correctoren de transparantie van het beoordelingswerk en de daaruit voortvloeiende feedback voor de kandidaten van hoger belang achtten dan de tijd die ze daarvoor nodig hadden. Dit blijkt ook uit hun algemeen hoge waardering voor de semi-automatische beoordelingsmethode.

De interbeoordelaarsbetrouwbaarheid van de correctoren werd vergeleken tussen blind versus zichtbaar beoordelen met 'checkbox grading' enerzijds, en met de traditionele beoordelingsmethode anderzijds. Bij het beoordelen met 'checkbox grading' kunnen de aankruisvakjes gekoppeld worden aan deeltcijfers. Door de aankruisvakjes aan te klikken die op de oplossing van de leerling van toepassing zijn, berekent de computer automatisch het totaalcijfer. Dit leidde tot het experimentele idee van blind beoordelen waarbij noch de deeltcijfers, noch het totaalcijfer aan de correctoren werden getoond. Uit de literatuur over rubrics is bekend dat beoordelaars durven afwijken van de criteria

wanneer hun holistische appreciatie van het werk van een leerling niet in overeenstemming is met de evaluatie die uit de rubric voortvloeit. Door de cijfers te verbergen, wilden we onderzoeken of we dit cognitieve conflict konden vermijden, wat zou moeten resulteren in een hogere interbeoordelaarsbetrouwbaarheid. Er bleek echter geen geschikte interbeoordelaarsbetrouwbaarheidsmaat te bestaan om deze onderzoeksvraag te beantwoorden. Er bestond namelijk geen voor toeval gecorrigeerde κ -coëfficiënt waarmee de betrouwbaarheid van meerdere beoordelaars, die voor elke leerlingoplossing één of meerdere aankruisvakjes selecteren, kon worden berekend. Bekende maten zoals de Cohen's kappa en Fleiss' kappa laten slechts de selectie van één item per leerlingoplossing toe. Dit leidde tot de ontdekking van een gegeneraliseerde Fleiss' kappa, die rekening houdt met alle informatie die 'checkbox grading' oplevert. Deze maat stelde ons in staat de onderzoeksvraag te beantwoorden: blind beoordelen verbetert de interbeoordelaarsbetrouwbaarheid wanneer beoordelingsschema's streng zijn en strikt moeten worden geïnterpreteerd, terwijl zichtbaar beoordelen beter geschikt is voor complexere beoordelingsschema's, omdat het zien van de cijfers de beoordelaars helpt de juistheid van hun eigen beoordelingswerk in te schatten. Vergeleken met de traditionele methode was 'checkbox grading' even betrouwbaar.

Om te onderzoeken hoe leerlingen reageren op de resulterende atomische feedback van 'checkbox grading', werd een vragenlijst afgenomen bij 36 van de 60 leerlingen die deelnamen aan het examen waarop dit onderzoek betrekking had. Vier van hen stemden in met semi-gestructureerde interviews. Leerlingen gaven de voorkeur aan de traditionele beoordelingsschema's boven 'checkbox grading' wanneer hen werd gevraagd feedbacksoorten te rangschikken van meer naar minder begrijpelijk. Toen leerlingen echter werden geïnterviewd met behulp van een think-aloud protocol, bleek dat ze 'checkbox grading' feedback gemakkelijker konden interpreteren. Bovendien waren 97% van de leerlingen het erop de vragenlijst mee eens dat de Examencommissie 'checkbox grading' zou moeten gebruiken als standaard beoordelingsmethode. Vooral de duidelijke link tussen de feedback en de totstandkoming van het cijfer werd hoog ingeschat. Hun begrip van dit soort feedback was gemiddeld hoog en kon niet gecorrigeerd worden met hun examencijfer. Dit betekent dat nagenoeg alle leerlingen, ook de minder goed presterende, de resulterende feedback vlot geïnterpreteerd krijgen.

De twee semi-automatische beoordelingsmethoden tonen twee waardevolle manieren waarop computers en leerkrachten kunnen samenwerken bij het beoordelen en geven van feedback aan leerlingen in het wiskundeonderwijs. De eerste aanpak was gebouwd voor individuele leerkrachten, de tweede voor een groep correctoren. Uit beide studies blijkt dat het geven van feedback altijd meer werk vereist dan het louter aanduiden van fouten of het meedelen van een cijfer. Bovendien lijken semi-automatische hulpmiddelen leraren vaak te motiveren om nog meer werk te doen in plaats van tijd te besparen. We concluderen met nog enkele ideeën voor vervolgonderzoek. Een eerste prioriteit is het slimmer maken van het suggestiesysteem van de eerste semi-geautomatiseerde tool. Door de ideeën uit de literatuur rond aanbevelingssystemen te integreren, kan het selecteren van feedback items om te hergebruiken vlotter verlopen. Daarnaast kunnen semi-automatische beoordelingsmethoden gelinkt worden aan Bayesiaanse netwerken om scherper zicht te krijgen op het individuele leerproces van leerlingen. Een Bayesiaans netwerk is een probabilistisch, grafisch model dat de bekwaamheid van een leerling in kaart brengt. Ook het onderzoeken van semi-geautomatiseerd beoordelen in andere settings (zoals peer feedback) zijn vruchtbare grond voor vervolgonderzoek.

✓ APPENDICES

APPENDIX A TEST ON LINEAR EQUATIONS (INCLUDING SOLUTION KEY)

Question 1

Solve the following equation.

$$\begin{aligned}\frac{3(x-1)}{5} - \frac{2(1-4x)}{7} &= x + \frac{x+1}{5} \\ \frac{3x-3}{5} - \frac{2-8x}{7} &= x + \frac{x+1}{5} \\ \frac{3x}{5} - \frac{3}{5} - \frac{2}{7} + \frac{8}{7}x &= x + \frac{x}{5} + \frac{1}{5} \\ \frac{21x}{35} - \frac{21}{35} - \frac{10}{35} + \frac{40}{35}x &= \frac{35}{35}x + \frac{7x}{35} + \frac{7}{35} \\ 21x + 40x - 35x - 7x &= 7 + 21 + 10 \\ 19x &= 38 \\ x &= \frac{38}{19} \\ x &= 2\end{aligned}$$

$\text{Soln} = \{2\}$

Question 2

Manipulate the formula to h .

$$\begin{aligned}
 A &= 2\pi rh + 2\pi r^2 && \text{naar } h \\
 &\Downarrow \\
 A - 2\pi r^2 &= 2\pi rh \\
 &\Downarrow \\
 \frac{A - 2\pi r^2}{2\pi r} &= h
 \end{aligned}$$

Question 3

Solve the following problem: The Junior Mathematical Olympiad consists of 30 multiple-choice questions. You receive 5 points for each correct answer. Each wrong answer obvious results in 0 points, but you get 1 point for each empty question. In this way, Jurgen got a score of 102 points with 4 wrong answers. How many answers were correct?

1) Keuze v/d onbekende

$$x = \# \text{ juiste antw.}$$

$$26 - x = \# \text{ niet beantw. vragen}$$

2) Vgl. ginst. en glemen

$$5x + 1(26 - x) + 0 \cdot 4 = 102$$

$$5x + 26 - x = 102$$

$$4x = 102 - 26$$

$$4x = 76$$

$$x = \frac{76}{4}$$

$$x = 19$$

$$\text{Oplosv.} = \{19\}$$

$$\# \text{ juiste antw.} = 19$$

$$\# \text{ niet beantw. vn.} = 26 - 19 = 7$$

$$\# \text{ juiste antw.} = 4$$

Antw: 19 antw. waren correct.

APPENDIX B PILOT STUDY DESIGN OF THE CROSSOVER EXPERIMENT

At the beginning of the summer of 2020, we organised a pilot study involving nine teachers, in order to rehearse the experiment. Based on the results of this pilot study, several changes were made to the actual study design. In the pilot study, teachers received handwritten copies of the students' tasks and were asked to provide handwritten feedback. They used the Start/Stop buttons in Moodle only to register their time for each question. Several observations are worth mentioning. Because teachers wrote their feedback on paper, they often forgot to push the buttons on the computer (thus contaminating the data). Organization also presented a major challenge, given the large amount of paper involved and the fact that each teacher had a different distribution of tasks in this condition. Most notably, the investigation actually revealed more about whether teachers could write faster on the computer or on paper than it did about the added value of reusable feedback. In the actual experiment, therefore, we changed the PP condition, using only a plain text box (see [Figure 1.3d](#)). In addition, the computer algorithm to produce the distribution of tasks between the conditions required in the pilot study only that, overall, every task must have been assessed the same number of times across both conditions. In all other respects, the distribution was completely random, meaning that some teachers were presented with unequal conditions, with one consisting of many more 'good' tasks and the other consisting of many 'poor' ones. This unequal distribution of tasks created practical problems and introduced bias into the data, as the feedback was more straightforward (and thus faster) to produce in one condition than it was in the other. For this reason, we decided to use the grades assigned by the teachers in the pilot study to divide the 60 tasks equally into three categories: good, moderate, and poor. We enriched the distribution algorithm with the constraint that every condition needed 10 good tasks, 10 moderate tasks, and 10 poor tasks, thereby ensuring greater equality among the conditions in the actual study.

APPENDIX C TEACHERS' SURVEY ITEMS BASED ON TAM (CHAPTER 1)

Scale	Item	M ± SD
Perceived Usefulness	SA is useful for me	5.22 ± 1.17
	SA can improve my performance as a teacher	5.19 ± 1.17
	SA allows me to perform my duties as a teacher more quickly	4.86 ± 1.27
Perceived Ease of Use	Using SA is easy for me	5.22 ± 1.10
	I find it is easy to let SA do what I want	4.47 ± 1.40
	The interaction with the SA system is clear and comprehensible	5.11 ± 1.12
Attitude Towards Using	Using SA is a wise idea	5.42 ± 0.91
	I like working with SA	5.14 ± 1.05
	It is a good idea to use SA	5.44 ± 0.91
Behavioral Intention to Use	I am planning to use SA in the future	5.17 ± 1.30
	I predict using SA in the future	5.19 ± 1.31
	I plan using SA in the future	5.08 ± 1.34

APPENDIX D

CODEBOOK OF ATOMIC FEEDBACK ITEMS

Atomic feedback items

An item is considered atomic when:

#	Guideline	Examples	When added
1	The item considers only one comment/mistake.	<ul style="list-style-type: none"> You subtract both sides with 26 instead of dividing both sides by 26. The method you used is correct. 	By definition
2	The item cannot be divided into meaningful hypothetical sub-items within the context of the question (e.g., when an item could hypothetically be divided into sub-items, but those hypothetical sub-items would always occur together, the item is considered atomic, as the division is not meaningful).	<ul style="list-style-type: none"> Convert $-1/7$ to denominator 35? → <i>This item cannot be divided in sub-items.</i> You do not take the wrong answers into account in your equation. There are 4 of them. → <i>It is given that there are 4 wrong answers in the word problem. It is unlikely that forgetting to include the wrong answers and not using this fact would occur independently.</i> 	By definition, refined after Iteration 2
3	The item acts as a 'chapter' to structure the feedback.	<ul style="list-style-type: none"> Identifying the unknown quantity and translation of the problem into an equation. 	Refined after Iteration 3
4	When addressing a structural error/misconception, the general error is separated from the specific mistake.	<ul style="list-style-type: none"> Error when adding fractions with unlike denominators. – $8x/7$ is not $4x/3$ 	By definition
5	The item contains a short subtitle (e.g., tip, notation, establishing equation, sign error) and an atomic remark/mistake.	<ul style="list-style-type: none"> Sign error: minus times minus is plus → <i>Dividing this into '- Sign error' and '-minus times minus is plus' would be possible, but it would have no real added value.</i> 	Iteration 1

6	The items concisely state that something went well/wrong at a certain location. (see Guideline 2 for being non-atomic)	<ul style="list-style-type: none"> • Step 1: correct. <p>→ <i>Dividing this into '- Step 1' and '-Correct.' would be possible, but it would have no real added value.</i></p>	Iteration 2
7	The hypothetical division into sub-items would give another impression or change the tone of the feedback.	<ul style="list-style-type: none"> • Final answer is correct due to a combination of errors. <p>→ <i>Dividing this into '- Final answer correct' and '-Combination of errors' would give a better impression than the teacher probably intended.</i></p>	Iteration 3

Non-atomic feedback items

An item is considered *non-atomic* if it can be meaningfully divided into hypothetical sub-items or if it violates the definition. In the task on linear equations, we distinguished the following cases in which this can occur:

#	Violations	Examples	Atomic alternative: division in hypothetical subitems	When added
1	The item discusses multiple errors/issues/remarks.	<ul style="list-style-type: none"> • Neither the choice of the unknown nor the initial equation is correct. 	<ul style="list-style-type: none"> • Choice of the unknown: wrong. • Start of the equation: wrong. 	By definition
2	The item contains references to both a structural error/misconception and the specific mistake.	<ul style="list-style-type: none"> • Improperly added fractions with unlike denominators, $8x7$ is not $4x/35$. 	<ul style="list-style-type: none"> • Error when adding fractions with unlike denominators. <ul style="list-style-type: none"> – $8x/7$ is not $4x/35$. 	By definition
3	The item contains both a comment/mistake and the location where the comment/mistake occurred.	<ul style="list-style-type: none"> • Step 1: the minus sign is written incorrectly before the fraction. 	<ul style="list-style-type: none"> • Step 1 <ul style="list-style-type: none"> – The minus sign is written incorrectly before the fraction. 	Iteration 1

4	The item makes an avoidable reference to another item.	<ul style="list-style-type: none"> • Bring 2 outside the parentheses first. • Idem to above comment for $\pi \cdot r$ 	<ul style="list-style-type: none"> • Bring outside the parentheses first. <ul style="list-style-type: none"> - 2 - $\pi \cdot r$ 	Iteration 2
5	The item makes links between solution steps.	<ul style="list-style-type: none"> • Step 2 is correct starting from the error in Step 1 	<ul style="list-style-type: none"> • Step 1 <ul style="list-style-type: none"> - Error: ... • I counted along with this mistake • Step 2 <ul style="list-style-type: none"> - Ok 	Iteration 2
6	The item contrasts an error/remark to the entire solution process.	<ul style="list-style-type: none"> • Well done, but there is no need to write a solution set here. 	<ul style="list-style-type: none"> • Well done. <ul style="list-style-type: none"> - You do not need to write a solution set in a word problem. 	Iteration 3
7	The item contains a reference to the number of times a mistake occurred.	<ul style="list-style-type: none"> • Adding fractions with unlike denominators goes wrong twice: <ul style="list-style-type: none"> - $3 \times 7 = 21$ instead of 15 - $1x = 35x/35$, not $35/35x$ 	<ul style="list-style-type: none"> • Adding fractions with unlike denominators: <ul style="list-style-type: none"> - $3 \times 7 = 21$ instead of 15 - $1x = 35x/35$, not $35/35x$ 	Iteration 3

APPENDIX E

MATHEMATICS EXAM OF THE FLEMISH EXAM COMMISSION

Part 1: Algebra - Complex numbers and matrices

Task 1 (2.5 points)

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.

🔍 First check-up

- No intermediate steps provided **max: 0.0**
- Solved using the *polar form of complex numbers*, which is impossible without calculator. **max: 0.0**

! Checking the calculation

- Correct complex conjugate $1-3i$ in the numerator. **+0.5**
- If the complex conjugate in the numerator is miscalculated or not applied, the student's answer will deviate from the solution key. Therefore, it is necessary to check the student's calculation individually for the indicated items.
- Check individually: Correctly multiplied by the conjugate binomial in the denominator. **+0.5**
 - Denominator may also be calculated immediately (= 29)
 - $\cdot(2-5i)$ is also fine (denominator in this case = -29)
 - Also fine if more steps were used; eg. first $\cdot(2+5i)$, next $\cdot(21+20i)$
- Check individually: Correct calculation of the numerator with intermediate step **+0.5**
- Correct denominator (=29 of =-29) **+0.5**
- Correct final answer in $a+bi$ form **+0.5 if calculation is fully correct**

Solution key

$$\begin{aligned}\frac{\overline{1+3i}}{-2-5i} &= \frac{1-3i}{-2-5i} \\ &= \frac{(1-3i) \cdot (-2+5i)}{(-2-5i)(-2+5i)} \\ &= \frac{-2+5i+6i-15i^2}{4+25} = \frac{13+11i}{29} \\ &= \frac{13}{29} + \frac{11}{29}i\end{aligned}$$

Task 2 (2.5 points)

Let $z_1 = b \cdot (\cos \alpha + i \cdot \sin \alpha)$ and $z_2 = c \cdot (\cos \beta + i \cdot \sin \beta)$, with $b, c \in \mathbb{R}_0^+$. Calculate the following expressions and write the answer in polar form.

a) $-5 \cdot z_1$

- -5 is correctly converted to polar form **+0.5**
It is not required that the conversion is done in a separate intermediate step, can be combined with the following intermediate step.
- Modulus and argument are correct. **+1.0**
Modulus must be positive!
 - It was converted back to $-5b \cdot (\cos \alpha + i \cdot \sin \alpha)$ **-0.5**
 - It only states $-5b \cdot (\cos \alpha + i \cdot \sin \alpha)$ **max: 0.0**
- The brackets around the argument and/or around $\cos \dots + i \cdot \sin \dots$ are missing. **-0.5 if maximum**
Mistake can occur in subquestion (b) too, but affects the score only once.

Solution key

$$\begin{aligned}-5 &= 5 \cdot (\cos 180^\circ + i \cdot \sin 180^\circ) \Rightarrow \\ -5 \cdot z_1 &= 5b \cdot (\cos(\alpha + 180^\circ) + i \cdot \sin(\alpha + 180^\circ)) \\ \text{OR:} \\ -5 &= 5 \cdot (\cos \pi + i \cdot \sin \pi) \Rightarrow \\ -5 \cdot z_1 &= 5b \cdot (\cos(\alpha + \pi) + i \cdot \sin(\alpha + \pi))\end{aligned}$$

b) $z_1 \cdot z_2^3$

Modulus correct +0.5

Note: $bc^3 \cdot (\cos \alpha + i \sin \alpha) \cdot (\cos 3\beta + i \sin 3\beta)$ has a correct modulus!

Argument correct +0.5

The brackets around the argument and/or around $\cos \dots + i \cdot \sin \dots$ are missing. -0.5 if maximum
Mistake can occur in subquestion (a) too, but affects the score only once.

Solution key

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

Task 3 (3.5 points)

The number of employees per category and the corresponding monthly wage (in euro) are shown in the table below.

	Number of employees	Monthly wage
Production manager	7	3750
Head of Department	4	6200
Director	1	9750

Let A be the 1×3 -matrix representing the number of employees for each category.

Let B be the 3×1 -matrix representing the monthly wage per category.

a) Complete and calculate $C = B \cdot A$

A and B are correct +0.5

Correct calculation of $B \cdot A$ +1.0

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

At most, there are 3 elements in $B \cdot A$ wrong:
Check the student's calculation individually for the following subquestions.

The error stems from a typing mistake in the graphical calculator.

There are more than 3 elements in $B \cdot A$ wrong:

No points for the explanations in subquestion (b)

The error does not solely stem from a typing in the graphical calculator.

Solution key

$$A = \begin{bmatrix} 7 & 4 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 3750 \\ 6200 \\ 9750 \end{bmatrix}$$

$$C = B \cdot A = \begin{bmatrix} 26250 & 15000 & 3750 \\ 43400 & 24800 & 6200 \\ 68250 & 39000 & 9750 \end{bmatrix}$$

b) Explain the meaning of the following matrix elements.

If the matrix element has no meaning, explain why.

- C_{11}

Right explanation +0.5

The wording might be different, but must refer to all keywords

Solution key

C_{11} : The total amount that is paid out monthly to the production managers as gross wage.

- C_{12}

Right explanation +0.5

The wording might be different, but has to refer to the key issues: eg., wage and category are different things, wrong categories are multiplied...

Solution key

C_{12} : No meaning - The wage of production managers is multiplied by the number of department heads.

- c) Give the matrix formula to model the following:
All the company's employees get a wage increase of 1,3%.

Formula is correct +1.0
B might be written out

Correct formula, but wrong calculation (so, not 3798,75; 6280,6; 9876,75) -0.5

Solution key

$$1,013 \cdot B \quad \text{OF} \quad \left(\begin{array}{ccc|c} 1,013 & 0 & 0 & \\ 0 & 1,013 & 0 & B \\ 0 & 0 & 1,013 & \end{array} \right)$$

Task 4 (3.5 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 extended coefficient matrix.

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

Solution key

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

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

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

Solution key

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

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

Part 2: Solid geometry

Task 5 (1.5 points)

- a) Find a set of parametric equations for the plane $\alpha \leftrightarrow 3x - 2y - 11 = 0$.

Set of parametric equations is correct +1.5
Attention: other possible solutions exists (other point and/or other direction vectors)

$\alpha \leftrightarrow$ is missing.

The curly bracket {} is missing. -0.5

$k, l \in \mathbb{R}$ is missing. -0.5

Solution key

$$\alpha \leftrightarrow \begin{cases} x = \frac{11}{3} + \frac{2}{3}k \\ y = k \\ z = l \end{cases} \quad k, l \in \mathbb{R}$$

Task 6 (1 point)

- a) Give a set of cartesian equations of the line a through $A(1, -2, 6)$ and parallel with the z -axis.

Set of cartesian equations is correct +1.0

$a \leftrightarrow$ is missing.

The curly bracket {} is missing. -0.5

The student's answer also contains a set of parametric equations. It is unclear which set (cartesian or parametric) is meant as the answer to the question. no points

Solution key

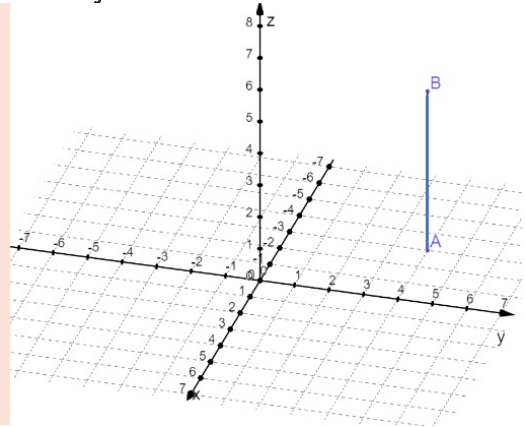
$$a \leftrightarrow \begin{cases} x = 1 \\ y = -2 \end{cases}$$

Task 7 (2.5 points)

a) Draw the line segment $[AB]$ with $\text{co}(A) = (-3, 4, 0)$ and $\text{co}(B) = (-3, 4, 5)$ in the given axis system.

- Point A is drawn correctly +1.0
- Point B is drawn correctly +1.0
Pay attention to the correct height!
- Can only be assessed when A and B are correct.
- Line segment AB is drawn correctly +0.5 if A and B are drawn correctly
 - Line AB is drawn -0.5 if A, B and $[AB]$ are drawn correctly

Solution key



Task 8 (4.5 points)

Calculate the distance between point $P(-8, 4, -1)$ and line $b \leftrightarrow \begin{cases} x = 4 + 4r \\ y = -11 + 7r, r \in \mathbb{R} \\ z = 5 - 4r \end{cases}$.

Solution key

Step 1: Searching the perpendicular projection P' of P on b

Method 1: by using the perpendicular plane

Perpendicular plane β intersecting P on b :

$$\beta \leftrightarrow 4x + 7y - 4z + t = 0$$

$$P \in \beta: 4 \cdot (-8) + 7 \cdot 4 - 4 \cdot (-1) + t = 0$$

$$\Rightarrow t = 0$$

$$\text{or: } \beta \leftrightarrow 4(x + 8) + 7(y - 4) - 4(z + 1) = 0$$

$$\Rightarrow \beta \leftrightarrow 4x + 7y - 4z = 0$$

Intersection point P' of b and β :

$$4 \cdot (4 + 4r) + 7 \cdot (-11 + 7r) - 4 \cdot (5 - 4r) = 0 (*)$$

$$16 + 16r - 77 + 49r - 20 + 16r = 0$$

$$81r - 81 = 0$$

$$r = 1$$

$$\Rightarrow \text{co}(P') = (8, -4, 1)$$

Method 2: by using a moving point on b

Point P' on b :

$$\text{co}(P') = (4 + 4r, -11 + 7r, 5 - 4r)$$

$$\Rightarrow \overrightarrow{PP'} = (4r + 12, -15 + 7r, 6 - 4r)$$

$\overrightarrow{PP'} \perp b$:

$$4 \cdot (4r + 12) + 7 \cdot (-15 + 7r) - 4 \cdot (6 - 4r) = 0 (**)$$

$$16r + 48 - 105 + 49r - 24 + 16r = 0$$

$$81r = 81$$

$$r = 1$$

$$\Rightarrow \text{co}(P') = (8, -4, 1)$$

Step 2: Determining the distance

$$d(P, b) = |\overrightarrow{PP'}| = \sqrt{(8 + 8)^2 + (-4 - 4)^2 + (1 + 1)^2} = \sqrt{16^2 + 8^2 + 2^2} = \sqrt{324} (= 18)$$

Step 1: Searching the perpendicular projection P' of P on b

Which method was used by the student?

- Method of the perpendicular plane
- Method of a moving point on b

Method of the perpendicular plane

- Correct general equation of the perpendicular plane β +0.5
- Correctly entering the coordinates of P in the equation of β +0.5
- Correct equation for β +0.5

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

- Student makes calculation- and/or sign errors: continue by checking the student's answer individually.
- Student makes other mistakes than calculation- and/or sign errors: no points for the rest of this question.
- Correctly filled in a random point on b in β (*). +0.5
- Correct calculation of r with at least 1 intermediate step +0.5
- Obtained correct coordinates of P' +0.5
Doesn't apply when β was wrong.

● Method of a moving point on b

- Determined the coordinates of a random point on b +0.5
- Correct calculation of $\overrightarrow{PP'}$ +1.0

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

- Student makes calculation errors in determining the coordinates of $\overrightarrow{PP'}$: continue by checking the student's answer individually.
- Student makes other mistakes than calculation errors: no points for the rest of this question.
- Correctly filled in a random point on b in β (**). +0.5
- Correct calculation of r with at least 1 intermediate step +0.5
- Obtained correct coordinates of P' +0.5
Doesn't apply when the coordinates of $\overrightarrow{PP'}$ were wrong

⦿ Step 2: Determining the distance

- Reasoning is correct: $d(P, b) = |PP'|$, with P' the perpendicular projection of P on b +0.5
- 1 intermediate step for the calculation of the distance +0.5
 - Correct calculation of the distance +0.5

Task 9 (2.5 points)

Is line $a \leftrightarrow \begin{cases} x + y - z = 0 \\ 5x - y - 2z = 0 \end{cases}$ parallel with line $b \leftrightarrow \frac{x-2}{4} = y + 3 = \frac{z-5}{2}$? Explain.

↑ Checking the distance vectors

- Correct direction vector of a with solution method included +1.0
 - Other multiples are also OK
 - Method must be included!
- Correct direction vector of b +0.5

● Explanation & Conclusion

- Can only be assessed when the direction vector of a and b are correct.
- Proper explanation +0.5
- Correct conclusion +0.5

Solution key

Direction vector (dv) of a ?

$$\begin{bmatrix} 1 & 1 & -1 & 0 \\ 5 & -1 & -2 & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & -0,5 & 0 \\ 0 & 1 & -0,5 & 0 \end{bmatrix}$$

$\Rightarrow dv(a) : (0,5;0,5;1) \sim (1,1,2)$

OR

$$\left(\begin{array}{c|c|c} 1 & -1 & 1 \\ -1 & -2 & 5 \end{array} \right) \sim \left(\begin{array}{c|c|c} 1 & -1 & 1 \\ 0 & -1 & 4 \end{array} \right) \sim \left(\begin{array}{c|c|c} 1 & -1 & 1 \\ 0 & -1 & 4 \end{array} \right) \sim \left(\begin{array}{c|c|c} 1 & 0 & 5 \\ 0 & -1 & 4 \end{array} \right) \sim \left(\begin{array}{c|c|c} 1 & 0 & 5 \\ 0 & 1 & -4 \end{array} \right)$$

$\Rightarrow dv(a) : (-3; -3; -6) \sim (1,1,2)$

$dv(b) : (4,1,2)$

$dv(a)$ and $dv(b)$ are no multiples (*)

\Rightarrow lines are not parallel

Part 3: Statistics & Probability

Task 10 (4 points)

For a particular study, they want to test 10 000 people for AIDS. Research has shown that 1% are carriers of the AIDS virus. In blood analysis, they combine the blood samples of 25 people and examine this mixture. If the result is negative, none of the 25 people is infected, saving 24 blood tests. If the result is positive, new blood tests have to be done for all 25 people. In this case, 26 blood tests have to be carried out.

Let X be the random variable expressing 'The number of blood analyses to be carried out for 25 people.'

a) Give a probability distribution for X . Probabilities should be rounded to 4 decimal places.

- Distinguished 2 possibilities (i.e., 1 and 26) AND correctly written down +0.5
 - Calculated probabilities are not important for this item
 - Applies when answer contains $P(X = 1)$ and $P(X = 26)$ OR table with distinction between X and x
- Correct result for $P(X = 1)$ +0.5
 - Intermediate step is not required
 - Also correct if calculated with the binomial distribution
- Not rounded or rounded incorrectly to 4 decimal places -0.5 once
0.7778, 77,78% and 77,7821% are all correct!
- Correct result for $P(X = 26)$ +1.0
 - Intermediate step is not required
 - Also correct if calculated with the binomial distribution
- Not rounded or rounded incorrectly to 4 decimal places -0.5 once
0.2222, 22,22% and 22,2179% are all correct!

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

- Student makes a right distinction between the two possible values for X , namely 1 and 26: continue by checking the student's answer individually.
- Student distinguishes other values than 1 and 26 for random variable X : no points for the rest of this question.

b) How many blood tests do you expect to have to carry out in a group of 25 people?

- Check individually: Correct intermediate step +0.5
 $E[X] =$ is not required to be written down.

- Check individually: Correct answer +0.5

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

- Student makes calculation error: continue by checking the student's answer individually
- Student makes other mistakes than a calculation error: no points for the rest of this question.

c) How many blood tests do you expect to save using this method in this study with 10 000 participants?

- Check individually: Correct calculation of the number of required blood tests. +0.5
- Check individually: Correct calculation of the saved number of blood tests. +0.5 if fully correct

Solution key

$$P(X = 1) = (0,99)^{25} = 0,7778$$

$$P(X = 26) = 1 - 0,7778 = 0,2222$$

OR

x	1	26
$P(X = x)$	0,7778	0,2222

Solution key

$$E[X] = 1 \cdot 0,7778 + 26 \cdot 0,2222 = 6,555$$

$$\Rightarrow 6,555 \text{ (or 6 or 7 blood tests)}$$

Solution key

$$10\,000 / 25 = 400$$

$$\Rightarrow 400 \cdot 6 = 2400 \text{ blood tests}$$

$$\Rightarrow \text{saving of 7600 blood tests}$$

$$\text{OR: } 400 \cdot 7 = 2800 \text{ blood tests}$$

$$\Rightarrow \text{saving of 7200 blood tests}$$

$$\text{OR: } 400 \cdot 6,555 = 2622 \text{ blood tests}$$

$$\Rightarrow \text{saving of 7378 blood tests}$$

APPENDIX F

ASSESSORS' SURVEY ITEMS BASED ON TAM (CHAPTER 4)

Scale	Item	Blind	Visible
		M ± SD	M ± SD
Perceived Usefulness	Checkbox grading is useful for me	4.3 ± 1.5	6.0 ± 0.9
	Checkbox grading can improve my performance as an assessor	4.7 ± 1.8	5.7 ± 0.8
	Checkbox grading allows me to perform my duties as a assessor more quickly	4.9 ± 1.9	5.4 ± 1.5
	Checkbox grading can make the assessment process at the Flemish Exam Commission easier	4.8 ± 1.5	6.2 ± 0.7
Perceived Ease of Use	Using checkbox grading is easy for me	5.1 ± 1.3	6.3 ± 0.6
	The interaction with the checkbox grading system is clear and comprehensible	4.5 ± 1.7	5.7 ± 0.9
	I find it is easy to let checkbox grading do what I want	3.9 ± 1.8	4.2 ± 1.8
	I find it easy to learn how to work with checkbox grading	5.7 ± 0.9	6.3 ± 0.6
Anxiety	Working with checkbox grading is a bit daunting for me	3.6 ± 1.8	2.5 ± 1.6
	I hesitate to use checkbox grading for fear of making mistakes I can't fix.	3.5 ± 1.8	2.0 ± 0.8
	I fear making mistakes I can't fix when using checkbox grading.	3.7 ± 1.9	2.9 ± 1.8
Attitude Towards Using	Using checkbox grading is a wise idea	4.6 ± 1.7	6.2 ± 0.7
	Using checkbox grading is a good idea	4.3 ± 1.9	6.0 ± 1.0
	I like working with checkbox grading	4.1 ± 1.9	6.0 ± 0.7
	It is a good idea to use checkbox grading	4.5 ± 1.7	6.1 ± 0.8
Behavioral Intention to Use	I am planning to use checkbox grading in the future	4.5 ± 1.6	5.5 ± 1.5
	I predict using checkbox grading in the future	4.7 ± 1.9	5.9 ± 0.9
	I plan using checkbox grading in the future	4.1 ± 2.0	5.4 ± 1.6

Feedback is the most powerful engine of any learning process. In mathematics education, the possibilities to assess automatically are thoroughly explored. However, students face difficulties expressing themselves mathematically on a computer and learning systems can often only assess the outcome, not the solving method. Research indicates that automated tests focus too much on procedural fluency at the expense of higher-order thinking questions. It takes much effort to develop digital tests, and teachers are sceptical of using automated assessments, meaning that paper-and-pencil tests still dominate mathematics classrooms. One of the characteristics of mathematical assessment is that wrong answers tend to exhibit patterns among the student population. Consequently, teachers often repeat their feedback and grades, bringing us to the idea of semi-automated feedback and assessment: by correcting handwritten tasks digitally, feedback can be saved and reused. This could lead to more elaborate feedback, time savings, and enhanced inter-rater reliability. Specifically, two semi-automatic assessment approaches were developed and studied.

In the first study, teachers write feedback for a student, and the computer saves it so that it can be reused when subsequent students make the same or similar mistakes. The concept of atomic feedback has been introduced to train teachers on how to write reusable feedback. Atomic feedback consists of a set of format requirements for mathematical feedback items, which has been shown to increase the reusability of feedback. A remarkable result was discovered during a crossover experiment with 45 mathematics teachers: the semi-automated approach led teachers to give significantly more feedback instead of saving time. Moreover, the teachers' feedback with the semi-automatic tool did not always have better properties than classic pen-and-paper feedback.

The second study was conducted in collaboration with the Flemish Exam Commission. Their traditional grading method of handwritten mathematics exams was transformed into a semi-automated one called 'checkbox grading.' Every assessor receives a list of checkboxes, and they must tick those that apply to the student's solution. Dependencies between these checkboxes can be set to ensure all assessors take the same path down the grading scheme. The system automatically calculates the grade and results in atomic feedback giving a detailed insight into what went wrong and how the grade was obtained. The approach requires more time for assessors and did not enhance inter-rater reliability compared to the traditional method (did not make it worse either). However, the resulting transparency and students' feedback were highly valued. Moreover, students could easily understand the resulting feedback, even the lower-performing ones.



Universiteit Antwerpen
ASOE | Antwerp
School of Education

ISBN 978-9-05726-797-8



9 789057 287978 >



Opening new horizons