A case-study of automated feedback assessment

Omid Mirmotahari, Yngvar Berg and Stein Gjessing
Department of Informatics
University of Oslo
Ester Fremstad and Crina Damsa
Department of Education
University of Oslo

***Abstract*—This study was conducted in a first year university course in computer programming. We investigated how the use of a digital assessment tool, consisting of quantitative weighting and which simultaneously generate qualitative formative feedback, improves the quality of the assessment and not least supports the students' further learning. Analysis of quantitative data generated by the digital system, as well as qualitative data from involved parties, shows clear improvements in validity and reliability in assessment. All students received automated feedback on their exam. In this paper, we show how the use if the tool contributed to students' learning and academical development. Furthermore, we present the process of developing this system for evaluation and automatic feedback, and conclude with results seen from teachers, reviewers and students' perspectives.**

Keywords: Grading, automatic feedback, first year course, programming subjects, object-oriented programming, higher education, criteria-based assessment.

I. Introduction

Studies have shown that getting feedback has a major positive impact on learning outcomes. Even though, over the last couple of decades, the number of students in higher education in increasing, we have not seen a corresponding change in focus when it comes to assessment. Rather, feedback practices appear to be largely seen as transfer of information controlled by the teacher. In (Tee & Ahmed, 2014), the authors point out that this is problematic because it ignores the way feedback contributes to students' self-understanding and motivation, emphasizing the importance of activating the student and using teacher assessment, students' self-assessment. With relatively simple steps early in the study, where the students receive feedback on their achievements, they can both get support for their own learning and be better at communicating their knowledge, skills and understanding at the exam. Feedback is important for developing meta-cognition and establishing good study habits and study and examination techniques. Therefore, it is important to address this, especially in the case of the novice students. We have therefore chosen to provide all first-year students with this individual feedback on how they performed the tasks and what they should work on, and we developed a digital tool that automatically generates feedback based on the assessment of handed in assignments.

Research suggests that a comprehensive approach to assessment and feedback based on socio-constructivist principles is what makes feedback the most productive in the sense that it supports student learning (Boud & Molloy, 2013; Esterhazy & Damşa, 2017; Juwah et al., 2004). This means that assessment and feedback are considered as part of an ongoing development process, where the student is involved as an active participant, and that summative and formative assessment are seen as part of the entire learning process (Rust, 2002). This also means that the students learn more from the exam, i.e the exam will not only be an assessment of learning, to determine the extent to which the students have reached the learning objectives, but also an assessment *for* learning, where the feedback helps a continuous learning in the subject. There has has been various attempts for feedback, including some on automatic feedback (Jiménez-gonzález et al., 2008; Malmi & Korhonen, 2004; Mirmotahari & Berg, 2018; Siddiqi, Harrison, & Siddiqi, 2010; Thelwall, 2000; Mirmotahari & Berg, 2017).

Reliability and validity are challenges that are important to exam and assessment. Studies show that reviewers or evaluators consider exam performance very different (Raaheim, 2000). In this context, the author (Raaheim, 2000) points out that the lack of criteria and review guidance is a key issue to explaining the lack of reliability. Central to the assessment tool developed and implemented in this study (Mirmotahari & Berg, 2018) is the development and use of clear criteria and guidance for reviewers. Much of the effort to establish this tool in regards to exams is mostly connected to the development of criteria, measuring scales and weighting. As students get insight into these criteria, it helps them understand what is required of them in an unprecedented academic setting, helping to create transparency and a non-threatening learning environment (Rust, 2002). Such transparency is one of the most important elements of constructive alignment (Biggs & Tang, 2007). Transparency, combined with automatic qualitative feedback, has potential to reduce the need for specific manual feedback, and also to ensure the students insight into what they have done correctly and what they have done wrong.

This study, based on the use of this tool in an introductory course (INF1010), illustrates object-oriented programming for first year students at the Department of Informatics at the University of Oslo and how the assessment and automatic feedback tool can be used. We will in this paper address the following with regard to automated feedback assessment:

(a) contribute to a more coherent relationship between assessment and teaching
(b) enhance validity and reliability of assessment
(c) reduce time spent for reviewing

Candidate number ->	
Assignment 1a	
Has interface	
Correct hierarchy	
Interface Administrator	
Calculated score (%)	0,00
Adjustment (%)	
Final score (%)	0,00
Final score in points	0
Text box	
Assignment 1b	
Correct decleration of all CLASS + Interface	
Constants are FINAL and given a value	
Correct constructor	
Method "ansvarskode()"	
"ansvarskode() instant return	
Calculated score (%)	0,00
Adjustment (%)	
Final score (%)	0,00
Final score in points	0
Text box	

Fig. 1: *Output from the program used by the reviewers. As it appears, for each sub-task there are different number of sub-goals and for each of these sub-goals the reviewer give 0-5 points (orange fields). The program then automatically calculates the total score for the given task. If the reviewer disagrees with the calculated score, they can use the field of adjustment.*

(d) generate data that provides the teacher with valuable information

The rest of this paper is structured as follows: in the next section, II, we provide a brief description of the self-developed assessment digital program. Then, in section III, we describe the research method for the study. In section IV we present various analysis of the collected data and the results with associated qualitative studies and experiences from teachers, reviewers and students. In section V we conclude this study.

II. Self-developed assessment program

The problem set for the exam in spring 2017 was developed in such a way that in each assignment the students could demonstrate how well they mastered one or more (largely related) learning goals. The entire exam set was developed to cover the maximum possible key learning goals and learning outcomes for the course. Mainly based on the workload, each of the 16 assignments received a weight so that the total

Assignment 1
Assignment1A (2 points)
1. Administrator (or similar name) should be an interface. This point have you answered very well.
2. Here you should display a proper subclass hierarchy with the names of the interface Administrator as well as the Employees, Nursing, Physicians, Occupational Therapists as well to the two classes that are both Chief and Nurses who can also administer. You have not answered this point so well.
3. The Administrator Interface shall be inherited by subclasses of Surveys and Nurses.
(a) The interface should preferably be drawn (higher on the drawing / sheet than) the classes as inherit it.
(b) It should be stated that it is an interface, preferably italics.
(c) Clear arrows will go up to the interface from the classes that implement it (two pieces).
(d) You get plus if you have shown that the superclass Employee is abstract.
These points have been answered very well.
In total, you have received 1.5 points on this assignment

Assignment 1B (9 points)
1. All classes (and the interface) must be declared correctly with extends and implements. You have answered this point well.
2. Constants should be declared final and get values in the designers. This The point has been answered very well.
3. The designers must call super (...) and this call should be the first in all constructors. This point you answered very well.
4. A string method, such as liability code (), should be found in the interface and implemented in the class of superior who can manage. This point have you answered very badly.
5. An instance variable must enter the liability code in this class. This point have you answered very well.
In total, you have scored 6.5 points on this assignment

Fig. 2: *Details of the textual feedback to a random student for assignments 1a and 1b.*

score yield 100 points. The weighting of the assignments was announced to the students in the assignment text on the exam. The examiner made a list of one to six learning goals for each assignment, which formed the basis of the review guidance and the feedback for the students. Based on the student's answer, the reviewer should assess how well the student had achieved these goals. Here a sixth graded scale was used where 0 = missing/absent, 1 = very weak, 2 = weak, 3 = fair, 4 = good and 5 = extremely good. The learning goals within each assignment were then weighted and the program calculated a score for the given assignment (rounded to the nearest half-point). The reviewer will see this score simultaneously and are therefore able to adjust the score in a separate post - adjustment - instantly.

The score per assignment (and the total score) was given to the candidates in the final feedback. The textual feedback for each assignment was obtained based on what they achieved on each sub or partial learning goal. The text phrases that were included in the feedback were based mainly on the supplementary review manual prepared by the examiner in advance of the exam. Both the actual sub-goals and the formulations used to describe these were chosen to so the feedback would be meaningful formative feedback to the students. As shown in Figure 2, the feedback contained a textual description of each sub-goal, followed by a textual description of the assessment. Initially, the report tool was created so that it was possible to generate different text phrases based on how a candidate actually achieved the sub-goals. The reason was that it was discovered that some sub-assignments gave rise to several correct answers, which did not coincide with the "blueprint" and thus failed to fit the phrases from the review guidance. The function seemed such that the reviewer could choose to asses one (and only one) of multiple assessment lines to indicate

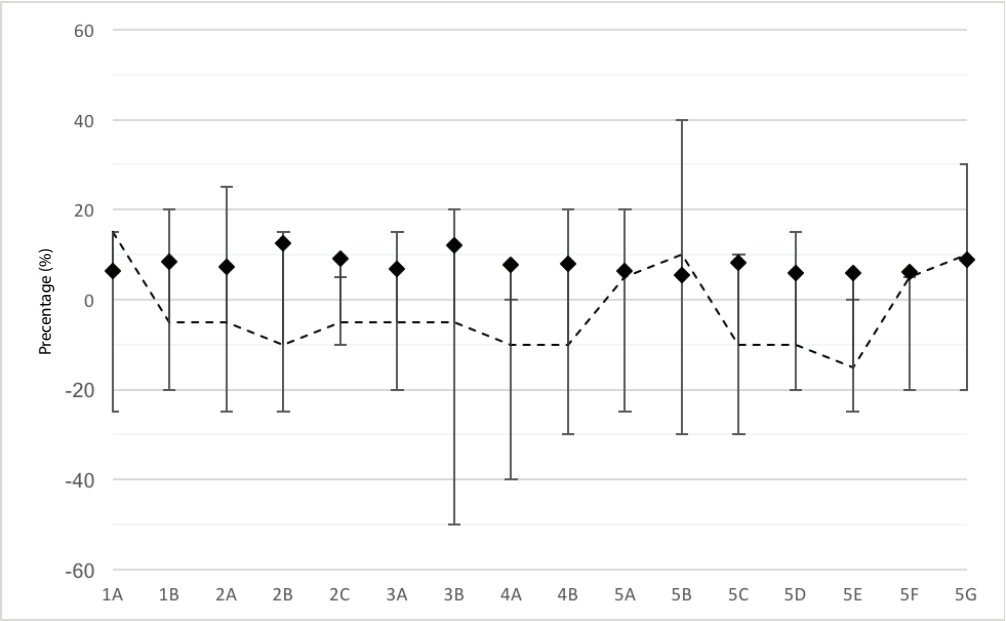

Fig. 3: *The relationship between all assignments on the exam and the reviewers score adjustment. The top and bottom lines represent respectively max. and min. percentage relative to the total score of the assignment. The dotted line represents the median of these adjustments. The diamond points show the number of responses that have been adjusted to show relative numbers.*

what solution the candidate had chosen. Although this was incorporated during the assessment itself and the data enabled the use of the function, we chose to skip this in the final feedback. Instead, the affected phrases were rewritten to better cover more interpretations of an assignment. The reason was that the descriptions still corresponded too badly with what the students had answered and that it might seem confusing if the students compared their feedback.

The students did not get the score for their achievement on the sub-goals, but rather a written feedback that represented those values. Nor was the internal weighting of the sub-goals within the sub-tasks presented to the students. For each sub learning goals, the reviewer had an opportunity to set no value, i.e. blank. Then the description of how well the candidate achieved this goal would not be included in the feedback. The reviewer also had the opportunity to enter additional comment in a text-box, which would be included in the final feedback to the students.

The report was generated by a proprietary, simple Python program that used a LaTeX template combined with a data model of the exam assignment. In this way, we could easily implement custom features such as hiding irrelevant feedback for missing answers. This functionality is surprisingly difficult to obtain by using predefined available solutions such as the Mail Merge function in Word, hence we developed specific software for this purpose. The students received their individual report directly on their university e-mail. The actual transmission of data between the assessment tool and the reporting tool meant some manual compilation of the data. By using this connection, we learned that problems may arise, particularly related to the text-boxes and the additionally comments made by the reviewer. Unfortunately, Excel 2016 still encode text strings with different character sets on, for example Mac and Windows, and the built-in data export feature does not necessarily allow cells in the worksheet to contain the same characters used as separate characters in the exported files.

III. Method

This study was conducted for 528 students who graduated in the subject INF1010 - Object Oriented Programming, spring 2017. The course is a compulsory subject in the second semester for all students throughout all study programs at the Department of Informatics at the University of Oslo. The course has been taught by the same teacher for the past 12 years and has in many ways kept the same content during this period. The teaching in the course extends over 14 weeks with four hours of lectures, two hours of group tuition and two hours of programming in lab with student assistants per week. Prior to the exam, all students must have passed 7 compulsory assignments. The final grade is only based on the final written 6-hour exam. The exam in 2017 consisted of 16 assignments of different weight, much corresponding to the previous years. There were 14 reviewers in addition to the teachers. Each answer was evaluated by at least two reviewers and a overlapping student mass between one to three review groups. The number of students enrolled for the course was 614, of which 528 were eligible to attend the exam in the spring of 2017. Data consists of data generated by the assessment tool, interview with the reviewers, and

questionnaire filled out by the students (20% response rate) after the exam and the automated feedback.
4).

IV. Results

One of the most discussed elements of such an automatic feedback system are the criteria. These criteria are formed to reflect the learning objectives of the subject by quantifying them. The extent to which such quantification succeeds is linked to factors such as differentiation (here used 0-5 points), the design of the assignment, the students' unique solution and the reviewers understanding of the sub-goals. One of the measures embedded in the assessment program is the ability to adjust each assignment. In Figure 3 we see the extent of adjustments made by all the reviewers. From the diamond points in the graph representing the percentage of all the responses that have been adjusted, we see that the majority of sub-tasks that have been adjusted are below 10% of a total of 528 responses. The number of adjustment for each assignment is around $\pm$ 20% of the task's relative points. In essence, assignment 3B and 5B deviate from these results, however, their median is in alignment within 10%. Upon closer evaluation of the two specific parts, it turns out that it is due to three specific answers that have had creative solutions that did not correspond the sub-goals that have been set. All of these candidates have also received feedback through the text-boxes during the assessment. If we read the median in Figure 3 we see that the trend is negative adjustment. This indicates that the reviewers want to give more points than they can because the low resolution of the criteria goes from 0 to 5. It is also evident from the interviews with the reviewers and their thoughts about this in section IV-A. Figure 5 shows the extent to which the reviewers have made use of the grading option for each sub goals. As expected, they used typically

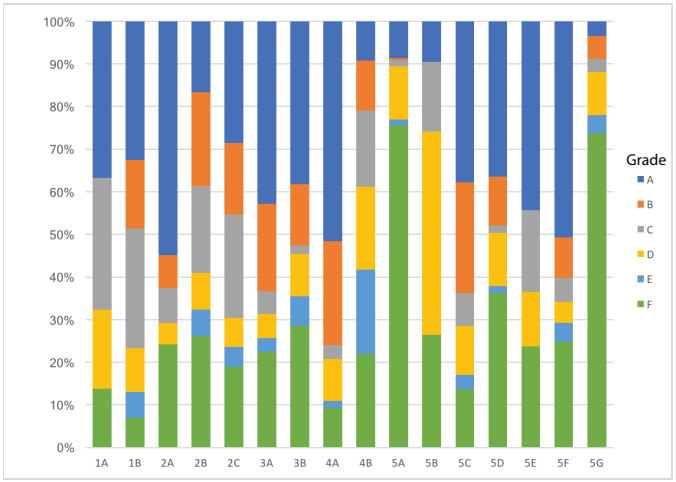

Fig. 4: *Displays the character distribution for the entire distribution over the assignments. This overview is a useful tool for the teacher to validate the exam assignments against both the review guidance and for determining the final grade.*

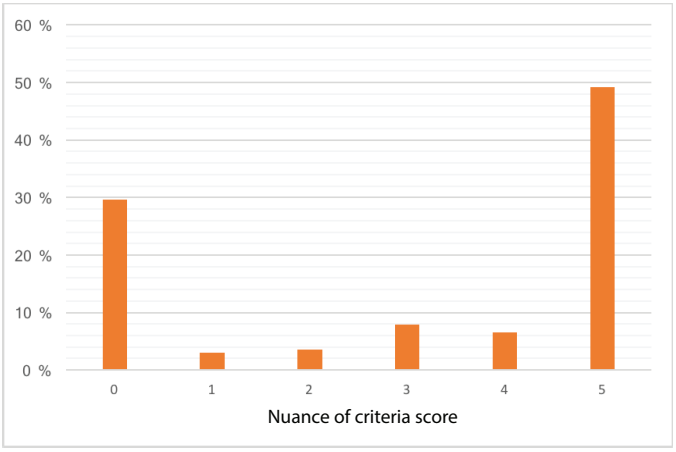

Fig. 5: This graph shows how the reviewers have made use of the 6 levels for the shading of the sub-goals. On the x-axis, the resolution is 0-5 points and the y-axis shows the relative value for each shade based on the accumulated sum of all sub-goals for all responses.

full points (5 points) when the sub goal was achieved and 0 points if not achieved. In Figure 6 we see the overview of the tasks the reviewers have used differentiated grades (1-4 points). This can also be seen in relation to which tasks have been most adjusted and which tasks have given the greatest nuances in terms of grades (Figure
All answers are randomly distributed and the number of reports per review varies from 6% to 22%. Figure 7 shows the individual grades of the reviewers based on the answers

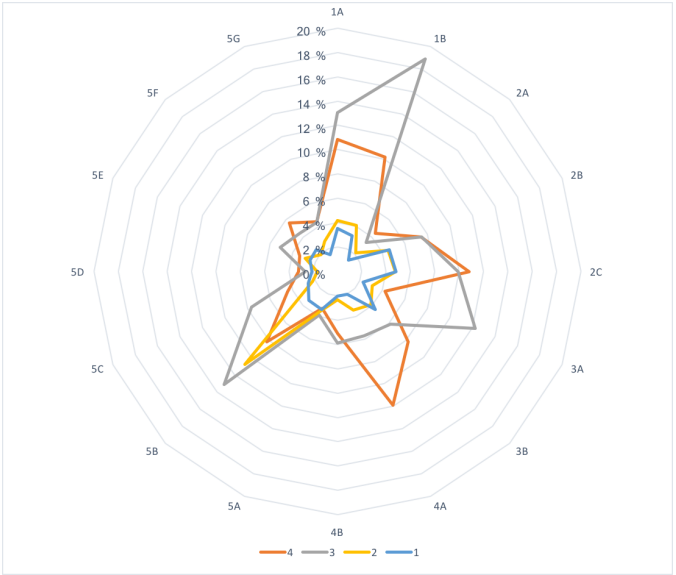

Fig. 6: The radar plot shows the extent to which the different nuance points are used, distributed on all tasks. In order to make this plot more readable, we have excluded the points 0 and 5. The places that are the lowest use of nuance are also the same as having the highest scores for points 0 and 5.

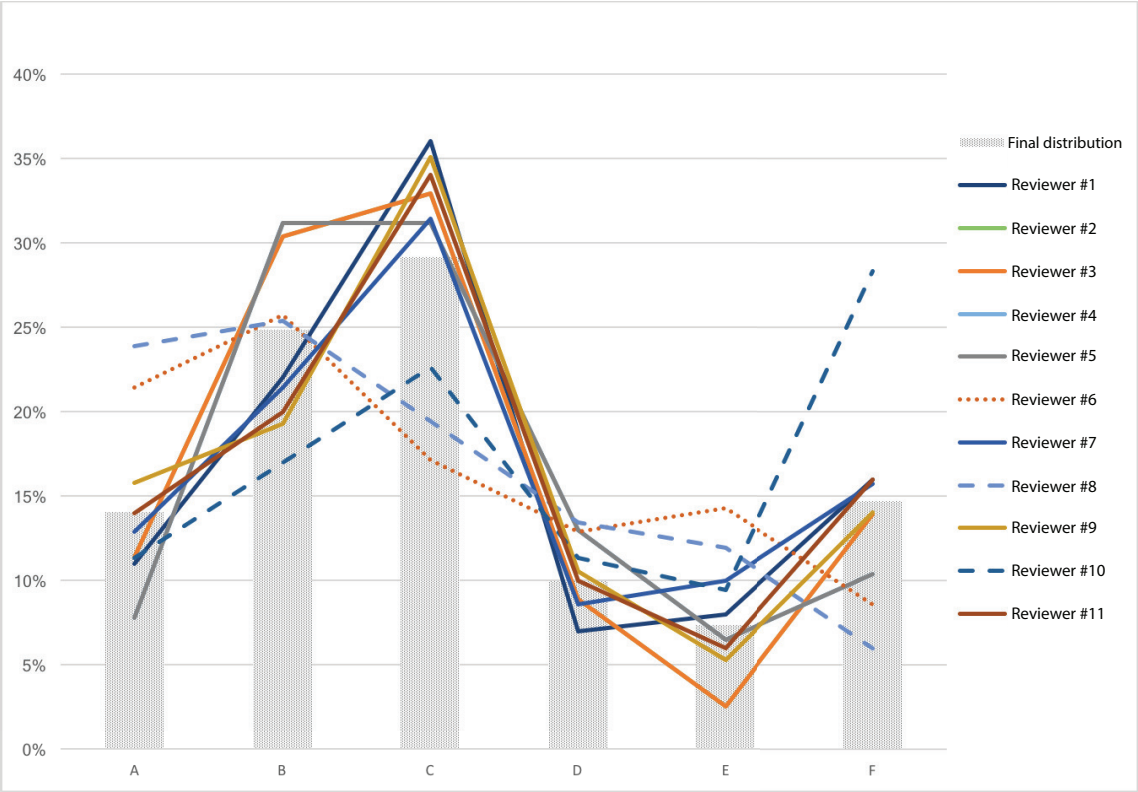

Fig. 7: *This graph shows the profile of the reviewers and grade distribution based on their selection of answers. The gray bars show the grade distribution of the final grade in the subject.*

they have been assessing, the selected reviewers are those who have assessed more than 10% of the student mass (of 528 exam hand-ins). The majority of the reviewers appear to have a normal distribution around a C+, corresponding to the entire grade distribution (gray bars). Nevertheless, three reviewers show a major deviation from this trend and are marked with dotted lines. It turns out that reviewer #10 is generally more strict than average, while reviewers #6 and #8 are less strict.

The teachers' main motivation has been to develop better exams, enhancing the reliability of the assessment, and provide feedback to the students in order to increase their learning outcomes. Secondarily, there are opportunities to avoid the reviewers having to assess several times for those students who makes a formal complain of their grade. Given all these upsides in this intervention, the assessment tool consumes a significant time, which can be described as first-time investment. The teacher has estimated an addition of approximately 10 work days as a result of this. In addition to these teachers, there have also been student assistants who have managed the program and data stream.

The teachers assessed 30 students as a control group to validate and test the assessment program. Their experience indicates that the learning goals that were written appeared more like a checklist and they have reason to believe that there will be a more fair and reliable assessment than with alternative methods. In particular, they bring forward those who have solved the task most similar to a blueprint and thus have scored full points. For the more creative (more or less good) solutions, they assume that using these points and their representative weight will be somewhat harder. It is therefore possible to use the adjustment on each sub-assignment as a balancing element. This adjustment was also used by the teachers.

Generating feedback to students also provides the opportunity to extract useful information to the teacher. In particular, it has been useful with the character distributed over the assignments, see Figure 4. With such an overview one can make data-driven decisions for weight between sub-goals, possibly eliminating goals that have been not been awarded correctly, seen on Figure 4. As an example, task 4A shows that over 75% of the students got top marks A and B, however should this task count as much as task 5B, which has 75% of the students obtained the grades E and F? If so, this will contribute to the reflection the teacher will do for the preparation of the next year's exam.

A. Experiences from the reviewers

Prior to the assessment, all the reviewers received an information letter explaining the procedure and how to use the assessment program. Before the actual assessment, some of the reviewers felt that the automatic feedback process would make the assessment more time consuming and one reviewer resigned because of this. The exam was on Friday, 16th of

June 2017. The reviewers received the student hand-ins late Saturday, 17th of June 2017 and the assessment process was completed on Wednesday, 28th of June 2017, one day before the formal deadline.

Following the intervention, both qualitative interviews and questionnaires have been conducted by all the reviewers. Two of the reviewers were first-time reviewers, while the third had reviewed earlier in this subject or in similar subjects. Unfortunately none of the reviewers have systematically done time tracking for the assessment part due to the fact that the reviewers receive a piece-rated payment and not an hourly payment. Nevertheless, the reviewers have estimated in retrospect how much time they used. The first evaluation took longer and all the reviewers think they have used between 45-60 minutes for the very first. Eventually, they have spent about 20-25 minutes per answer. Given the method, they have a good distribution between correcting horizontally or vertically. The usual assessment procedure is to review vertically, i.e to assess the same assignment for all students after each other before assessing the next assignment. Horizontal assessment means to assess a complete answer/exam for one candidate before assessing the next candidate. There are pros and cons of these two ways to do assessment. For vertically assessment it is argued that it gives a more fair assessment, while for horizontal assessment it is argued for a uniform impression for the given candidate which may give a more accurate assessment. Common to both review groups was the emphasis that sub-goals were a good way to obtain the most fair assessment. Some reviewers have used the text-boxes extensively, but eventually gave up for mainly two reasons. One is obviously the time investment - here they show that they did not quite understand the students would receive a automated feedback based on the score. The second reason is due to the critical feedback of the reviewers on the comments, pointing out that it quickly became a keyword in the text-boxes.

Concerning the adjustments made, the reviewers generally agreed that the learning goals were descriptive enough to make the assessment and therefore in particular cases the adjustment made were isolated and distinctive. For the vast majority of sub goals, it could have been good enough with three levels for grading. However and contrary, they pointed out that they would also like the possibility of having a six-level scale on all the goals in the future, even if they are not used to everyone. Elaborating on the adjustment types showed that there were no conclusive use in terms of adding or deducting points. Majority of the deduction of points were due to students lack of fundamental knowledge, to much unnecessary code and apparently just a transcript of lecture notes. There are pros and cons of using minus points, but we think that the adjustment allowed applying minus points indirectly and thus avoid the disadvantages minus points will have for weighting and complexity for calculating the final results. For the case of positive adjustment, the most frequent follow-up error was that the result of the sub-goals was lower than the students showed understanding and thus the reviewers chose to adjust/compensate. There was a general consensus about the use of the adjustment, all reviewers used a overall adjustment for the all the sub goals in an assignment rather than adjusting the sub goals score, which we regard as extremely positive for our collected data.

Given the inter-rate reliability, it appears that the review pairs for each exam hand-ins experienced surprisingly small differences, i.e relative difference in points. The reviewers expressed that the differences was exclusively in the grades. The reason lies in the quantification of sub-goals against static values for grade. This means that a 0.5-point difference between two reviewers can actually constitute an entire grade difference. The validity of the assessment has been verified by the fact that all the review pairs have gone through the deviations and issued a common unified score. Further validation is also done for all students in the grade threshold zone, defined by a teacher at $\pm$ 2 points. All those students have been carefully reviewed and after a comprehensive assessment from both reviewers, the final grade has been decided.

From section II we find that the sub goals are determined by what the teacher has thought about the solution for the assignment. These sub-goals have been a good guide for the reviewers, especially the first-time reviewers have considered them a better aid than the review guidance. The reviewers have also had the opportunity to discuss the details with each other and the teacher, but they did not use this offer to any extend.

B. Experiences from the students

More than 67% of students say that the automatic feedback matches very much with their own self-evaluation of their performance and that they have benefited greatly from the feedback, as quoted in Q:1.

> *«Automatic feedback was very useful to me and provided insight into my own capacity and areas for improvement. I would not ask for the reason for the grade, but learned a lot from the feedback. The feedback convinced me enough and in retrospect, I do not disagree with the grade. Automatic feedback should be part of all subjects where possible!»*
>
> *(*Student #2978171*)* Q:1

43% of the students say that the feedback has encouraged them to contact fellow students and that the feedback has been widely used in discussion. This illustrates that the students experience the feedback as useful for further learning and over 45% of students say they have read and used the feedback several times after the assessment and in further studies. The following quotes, Q:2 and Q:3, also show that the feedback and a transparent assessment gives the students increasing confidence in the assessment result. From an economic aspect, one formal complaint from a student is enough to ignite new review committees and many work hours administrating the process. From a university management perspective, any

small investment in prior to exams for prevent complains from students are very welcome.

> *«It was very reassuring to get automatic feedback, and I felt much safer on the assessment that had been done on the subject.»* (Student #2977287) — Q:2

> *«The scheme is good and should be continued. I missed one of the grade margins by 0.5p, but still felt that the feedback was clear enough and the score was well-founded that it would hardly cause a complaint. It also gave insight into what things I should have thought of or done better, which is very useful.»* (Student #2977450) — Q:3

> *«Auto feedback was a wonderful surprise. I wish all subjects did this. I got an answer to everything I was wondering about my grade - before I realized it was something I was wondering about. I also felt that the grade was more fair after receiving such a thorough and completed automatic feedback. This feedback is MUCH BETTER than the feedback I actually requested. Keep it up!»* (Student #3073951) — Q:4

In relation to the role the feedback has played, for whether students have chosen to complain or not, 54% of those who did not complain argued that the feedback provided insufficient insight into the assessment of their choice. While, 75% of those who complained based their complaint on the feedback. The complaint level for the whole course was 5%.

V. Discussion

As the results show, a good preparation is important in order to succeed with good criteria and the opportunity for these criteria. Even though we can claim that automatic feedback reduces evaluation time, the actual total time spent will be almost the same as before. Much of the time spent for the reviewers goes to the teacher who develop the review guidelines and the establishment of criteria. It can be argued that the time will also largely lead to a constructive alignment in the course and also contribute to increased quality of the questions. In line with Raaheim (Raaheim, 2000), this study shows that there is strong correlation between clear criteria and review guidance, and the reliability of the assessments. Experience from the reviewers shows high correlation, especially through the amount of adjustments they have made on the exam review, Figure 3. These adjustments have mainly been made for solutions that have been correct, but have been very different than the «blueprint». In the discussion for the next study, we will be able to look at the relationship between the resolution, here used 0-5 points, and the adjustment amount. We see from Figure 5 that the use of the extremes in the gradation scale is significant, but it can also be argued with the amount of sub-criteria. There were a total of 50 sub-goals and 528 responses, which gives 26,400 data points, then it will of course give a great impact if any of the sub-goals have been to discrete binary quantification. With Figure 6 we can see which sub goals have been best suited for quantification. This does not, however, mean that the tasks that do not appear here can not or should be quantified, but that there has not been a need for a scale of 0-5 points.

In relation to the validity of the assessment, we can extract the amount of adjustments that the reviewers have made. The adjustments can also be viewed as the biased assessment and directly the discretionary assessment the reviewers make. This is seen in conjunction with the reviewers profile (Figure 7) we can extract and to some extent calculate the deviation from the others. Here it can eventually be automated so that the reviewers' assessments are normalized in relation to each other and achieve even higher reliability. One aspect of validity in the assessment is the initial calibration of reviewer in relation to the teacher. Through this study, we have, based on the teachers' and the reviewers feedback, understood that the sub-goals with the sub-criteria have in many ways contributed to a sort of checklist for what to look for in an assignment. As this list has been so detailed and the ability to enter a score (0-5p) there has been no request for further information meetings or writing with reviewers. In many ways, the reviewers have self-calibrated through using this program. The subject's grade distribution for the year's exam follows the close distribution of the subject, and as Figure 4 shows that the majority of the reviewers also follow the same distribution, which in turn reinforces the impression that the program strengthens reliability and validity.

VI. Conclusion

In this paper we have presented a study done for a major informatics topics, n> 500 students, with automated feedback to the students based on their exam hand-ins. As the results show, a good preparation is important in order to succeed with good criteria and their grading. Even though we can claim that automated feedback reduces reviewers time, the actual total hourly budget will be almost the same as before. Much of the time saved for the reviewers goes to the teachers who invests in the development of review guidance and the establishment of criteria. As Raaheim also highlights (Raaheim, 2000), this study shows that there is a strong correlation between clear criteria and review guidance, and the reliability of the assessments. Experience from the reviewers shows high inter-rate reliability, especially through the amount of adjustments they have made to the exam. These adjustments have mainly been made for solutions that have been correct, but that have been very different from the "blueprint". In the discussion for the next study, we will be able to look at the relationship between the nuance range, here used 0-5 points, and the adjustment

amount. Students' feedback indicates that the automatic qualitative feedback is perceived as positive both in the sense that they reinforce the understanding and trust of the assessment behind the final grade and that the feedback is perceived as contributing to their professional development. Both the understanding of what they have received and where they have failed, and comments that give a direction for further work are perceived as valuable. This corresponds to the literature on formative assessment, highlighting the meaning of 'feedforward' (Nicol & MacFarlane-Dick, 2006). The feedback also indicates that the students take the comments active for use in further professional work, not least in cooperation with fellow students. Thus, this form of feedback seems to support not just academic learning but also the students' meta-cognition and self-regulation - competences essential for success in higher education (Bransford, Brown, & Cocking, 2000).

VII. Further work

Although this case-study has been for a programming course, the framework can easily be adapted to other field and courses. Throughout our work with this case we are under the impression that the main contribution of working with automated feedback assessment comes from the dialog between the teachers. The discussion on how the learning goals of the course can and will be assessed as well as whether the students achieve the expected learning outcomes through the assignments.

References

Biggs, J., & Tang, C. (2007). *Teaching for Quality Learning at University Third Edition Teaching for Quality Learning at University* (Vol. 3th edition (1th edition 1999)). Open University Press. doi: 10.1016/j.ctcp.2007.09.003

Boud, D., & Molloy, E. (2013). *Feedback in Higher and Professional Education: Understanding it and doing it well*. Routledge.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How People Learn*. doi: 10.17226/9853

Esterhazy, R., & Damşa, C. (2017). Unpacking the feedback process: an analysis of undergraduate students' interactional meaning-making of feedback comments. *Studies in Higher Education*, *5079*, 1–15. doi: 10.1080/03075079.2017.1359249

Jiménez-gonzález, D., Álvarez, C., López, D., Parcerisa, J.-m., Alonso, J., Pérez, C., … Tubella, J. (2008). Work in Progress – Improving Feedback Using an Automatic Assessment Tool. *ASEE/IEEE Frontiers in Education Conference*, 9–10.

Juwah, C., Macfarlane-dick, D., Matthew, B., Nicol, D., Ross, D., & Smith, B. (2004). Enhancing student learning through effective formative feedback. *The Higher Education Academy Generic Centre Enhancing*, *1*(68), 1–41.

Malmi, L., & Korhonen, A. (2004). Automatic feedback and resubmissions as learning aid. *Proceedings - IEEE International Conference on Advanced Learning Technologies, ICALT 2004*, 186–190. doi: 10.1109/ICALT.2004.1357400

Mirmotahari, O., & Berg, Y. (2017). Individuell "automagisk" tilbakemelding på skriftlig eksamen. *Nordic Journal of STEM Education*, *1*(1), 287–293.

Mirmotahari, O., & Berg, Y. (2018). Structured peer review using a custom assessment program for electrical engineering students. *IEEE Global Engineering Education Conference, EDUCON*, *1*. doi: 10.1109/EDUCON.2018.8363339

Nicol, D., & MacFarlane-Dick, D. (2006). Formative assessment and selfregulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. doi: 10.1080/03075070600572090

Raaheim, A. (2000). En studie av inter-bedømmer reliabilitet ved eksamen på psykologi grunnfag. *Tidskrift for Norsk Psykologiforening*, *37*, 203–213.

Rust, C. (2002). The Impact of Assessment on Student Learning: How Can the Research Literature Practically Help to Inform the Development of Departmental Assessment Strategies and Learner-Centred Assessment Practices? *Active Learning in Higher Education*, *3*(2), 145–158. doi: 10.1177/1469787402003002004

Siddiqi, R., Harrison, C. J., & Siddiqi, R. (2010). Improving teaching and learning through automated short-answer marking. *IEEE Transactions on Learning Technologies*, *3*(3), 237–249. doi: 10.1109/TLT.2010.4

Tee, D. D., & Ahmed, P. K. (2014). 360 degree feedback: An integrative framework for learning and assessment. *Teaching in Higher Education*, *19*(6), 579–591. doi: 10.1080/13562517.2014.901961

Thelwall, M. (2000). Computer-based assessment: a versatile educational tool. *Computers & Education*, *34*(1), 37–49. doi: 10.1016/S0360-1315(99)00037-8