Innovating assessment practices using automated feedback in software in computer science education

Omid Mirmotahari, Yngvar Berg
Department of Informatics, University of Oslo, Norway
Crina Damşa
Department of Education, University of Oslo, Norway
omidmi@ifi.uio.no, yngvarb@ifi.uio.no, crina.damsa@iped.uio.no

Abstract: This paper presents an innovation that includes modalities of generating and providing automated formative feedback to Computer Science undergraduate students. The paper addresses a challenge posed by the increasing numbers of students and amount of information teachers must deal with. A software solution providing automated formative feedback has been developed and used in this study in order to both support the students learning process and to use a criteria-based approach for assessment and peer-feedback. The automated data allowed for employing an algorithms-based analysis. Such analyses and findings provided the lecturer with input for new ways and tools to improve the students' learning gains, while the logs from the software program can further be used through machine learning and AI to teach the program able to self-determine the grade.

Context of the innovation work

This paper presents the development and use of a software program and criteria developed for providing automatic feedback in a Computer Sciences undergraduate program. A user experience study was conducted to identify the students' experiences both with the use of the system and with receiving feedback in this manner.

Feedback is viewed as a pedagogical strategy in teaching-learning environments that has potential to facilitate the students' learning in a meaningful manner (Jansson, 2006). Generally, evidence supports the idea that feedback to students plays a significant role for student learning (Sadler, 2010). Many studies have focused on the effects feedback has on individual students' learning outcomes and study achievements (e.g. Zimbardi et al., 2016). Greenhow's (2015) argues for the inclusion of a formative component when developing digital environments for providing feedback. Her study shows that only providing the students with a summative evaluation (grades or point sum) does not stimulate deep engagement with the knowledge or following learning activities. With an increase in student population and in the amount of knowledge that needs to be conveyed through the curriculum-related activities makes the task of organizing and providing feedback and assessment that has a formative value becomes even more difficult.

Tools for generating automated feedback have been developed in the last years, but there is a new bust in the field, with technologies of increased sophistication providing more realistic solutions. Common for all these applications are that answers can be expressed in a formal way, which makes it feasible to automatically analyze them. Automatic assessment of answers composed of free textual sentences, however, is evidently much harder. Usually, the automatic feedback in an assessment system is given by a short indication of whether the answer is wrong or right, some systems also allow a resubmission. Malmi and Korhonen (2004) conducted a study that allows the students to resubmit assignments, but the exercise randomly changed the initial data, thus giving a positive effect on the learning outcome. Lewis and Sewell (2007) attempted to examine the effectiveness of providing formative feedback for summative computer-aided assessment, by giving individualized feedback derived from each of the five results sections of the assessment was provided to each student. A study by Siddiki et al. (2010) examined how an automated short-answer marking system can be effectively used to improve teaching and learning at university level, finding that such a system can provide practice tests and immediate feedback to students regardless of the size of the class. But the system did not allow features that can provide detailed statistical analysis of students' performances for both lecturers and students so that each may adjust or modify their teaching or learning approach for the course. This brief review illustrates an issue in the field, namely, that the combination of structured, individualized assessment and automated feedback is still not widely examined. While technologies are developing at a fast pace, the research and empirical evidence of the pedagogical solutions such technologies are providing is limited and still marginal to the learning field and educational practice.

Our innovation addresses these issues and this study aims to provide a better understanding of: how a software program developed for providing automatic formative feedback in a Computer Sciences undergraduate course can be implemented, how the automated feedback contributes to improve learning and how the students

experience both the use of the system and receiving feedback in this manner. Whether students engage productively with feedback, whether it enhances their learning and performance, and whether automated fee feedback can have meaningful role in these processes are questions that require later empirical examination.

Description of the intervention: The software program for individualized assessment and automated feedback

The software program (Mirmotahari, 2017, see Figure 1) was initially developed to facilitate exam assessment, but it gradually displayed great potential for being used to give formative feedback. The program was designed with several stakeholders and users in mind, namely (i) exam evaluators from whom this program is time-saving and provides detailed guidelines for assessment procedures; (ii) students, who will benefit an objective assessment of their exam assignments; and (iii) the teachers, for whom the program will provide a range of analyzes and reports on students' answers, level of knowledge and learning outcomes.

How does this software program work? The evaluators tick the checkboxes for which "mistakes" the student has made and each of these arguments have a weight which impacts the grading. The degree of impact for a given mistake is predefined by the teacher and closely linked to both the criteria and the Impression scale bar (see top-right box in Figure 1). Nevertheless, if the evaluator finds the impact of a certain mistake to be weighted too much, s/he can override the weight by manually adjusting the scrollbar at the Impression scale. The program will constantly follow and log all the choices made. All data is then processed so that the system can calibrate each evaluator 's discretionary assessments of the assignments. Through the quantified criteria, the assessment program will perform a variety of calculations. These calculations lead to an aggregation of several predefined text phrases that compile to a comprehensive textual feedback. This individual feedback is automatically generated based on different thresholds for the criteria determined in advance. Each criterion has an underlying text phrases and depending on the weighted accumulated value of the criterion, a specific text phrase will be picked out. Thus, all criteria are put together for a *comprehensive individual feedback*. The program can provide both the arguments for the grade as well as an individual formative feedback. If the evaluator or the teacher chooses to provide students with both arguments for their grade and an individual formative feedback, the first part of the feedback will consist of the arguments for each assignment and grade. The second part will be an individual formative feedback based on feedforward principles.

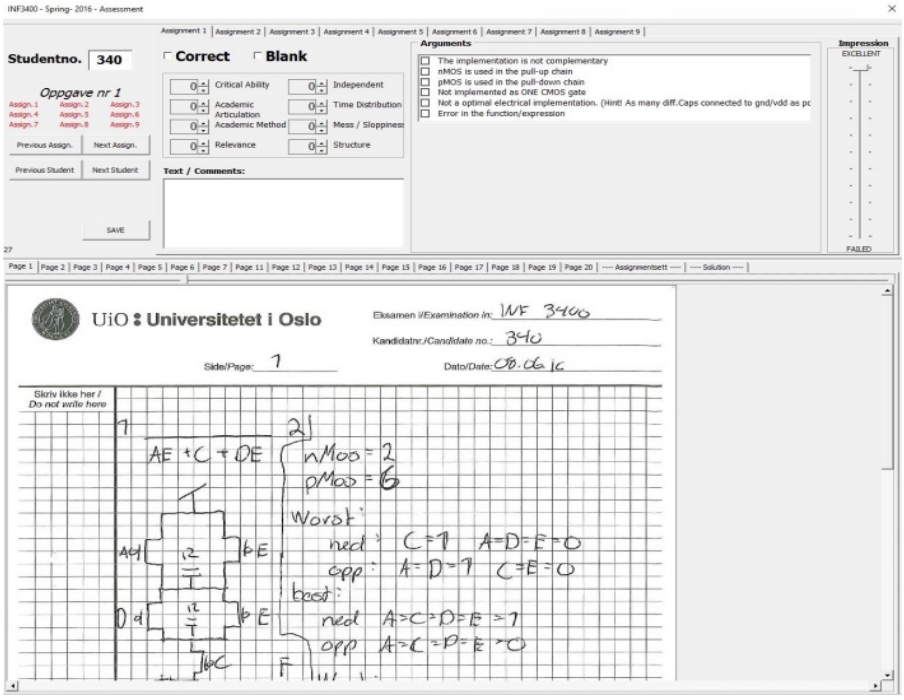

Figure 1. Screenshot of the software program interface

The user-interface has been designed primarily for the evaluator/teacher. Each assignment has its own tab at the top, but they all have some parts in common. The parts all assignments have in common are the blocks called Criteria, Arguments and Impression. For the Criteria, these are the main criteria which the teacher has predefined and are closely derived from the course taxonomy. For the Arguments, they are also predefined by the teacher. The arguments are "errors/mistakes" the students usually do. The Impression scale is a continuous scale with the only labeling of "Failed" – to – "Excellent", not divided by grade or points.

The formative assessment criteria

The main component of the assessment program is the generation of the criteria, their weight in measurement and the textual phrases linked together. A taxonomic model (inspired roughly by Bloom's taxonomy) was used to develop a set of criteria that focused on the students' learning understanding of abstract knowledge and the way to employ this in solving computing problems. The set of developed criteria that were implemented in the system and were attached particular types of feedback is presented below:

- Understanding (F) - the students' ability to demonstrate good understanding;
- Critical Abilities (K) – the capacity to be critical to their answers when asked to solve a task in different ways;
- Neatness (R) - structure and tidiness for the entire answer;
- Writing and Language (S) - mainly the clarity of the answer, using a professional language and terminology;
- Blank answers / missed assignments (O) - number of missed assignments in intervals of (0), (1-3), (4-6), (7->);
- General (G) - general points that are not captured by the above criteria. These can be grouped into:
    - Explanation of mind (FT) - assumptions or mistakes that make it difficult to follow the thought
    - Time for important tasks (TVO) - part of the most important tasks (the highest weighting) is not answered or that the student
    - Time allocation (TD)- use of the exam time
    - Exam Training (E) - unstructured and unprioritized solution to the weighted assignments
    - Time to Control (TK) – time management, finding alternative ways to solve the same task to see if the answers become the same
    - Presentation (P) - writing and language is not clear and comprehensive enough.

There is a trade-off for choosing the number of criteria. Too many criteria would make the assessment process long and with too many parameters which may lead to misunderstanding between the reviewers. On the other hand, too few criteria would restrict the nuances of assessing. The golden mean would be to have a hierarchical construction of criteria. On the top level, the main goals of the course, i.e. course taxonomy, can act as a grouping tag/name. Within each group there can be a few and specialized criteria, called sub-criteria. At this level of sub-criteria there can be different ways of putting in a weight, e.g. number scale, grades and so on. The criteria, sub-criteria and their representative weight generated a multidimensional matrix. This matrix is used to obtain a rich nuance for giving individual feedback for an assessment. As an example, one teacher had the following hierarchy shown in Figure 2. The teacher had chosen mostly to denote the weights as ´1´, ´0´ or ´-1´, meaning positive, neutral and negative impact, respectively. For the criterion "knowledge", the sub-criteria had a different weighting scheme. Under the "grade" the scale was A-F, while for the "Arguments" there were letter-tags corresponding to a "predicted" error.

Knowledge		Communication			Technical			
Grade	Arguments	Critical Thinking	Academic terminology	Reasoning	Writing	Orderly	Structure	Time distribution

Figure 2. An actual example for a criteria categorization

Any given assessment requires a set of rules, criteria or guidelines for the evaluator to assess a task. When a teacher authors an exam or an assignment s/he has a prior knowledge of what the students may fail on. These predictions of possible errors or mistakes are collected and tagged in the program. The list of possible errors and mistakes are shown to the evaluator when assessing the exam/assignment. It is this list that is displayed in the Arguments box in Figure 1.

Description of the intervention: Automated individual feedback

The back-end of the program is constantly monitoring and analyzing the evaluator's choice and overruns. The results of these analysis lead to individual feedback to each student. The feedback consists of three main parts; (i) academic feedback and discipline-based justification of grade, (ii) personal feedforward and finally (iii) a profiling for the learning outcome. The length of the feedback is entirely dependent on the amount of choices the evaluator has made and the accumulated sum of the weights of the criteria throughout the whole assignment. The accumulated sum for each criterion is normalized to the classes and based on predefined thresholds groups the results. Each criteria result will give a corresponding match in one cell in the matrix holding text phrases. The number of cells to be selected depends on the significance, i.e. the normalized sum for a criterion and formed as a sequence of text.

Dataset and analytic strategy

All the students' hand-in assignments were scanned and automatically uploaded into this assessment program. After each iteration, the students were asked to answer an online questionnaire. The average response rate has been 77%. Different questionnaires have been used to collect in answer regarding one or more of these topics: questions about the assignment; perception of the feedback received; evaluation of the technical aspects of the assessment program (computer program, usability, and time usage); development in relation to the professional domain; their experience of being a peer reviewer (evaluator); learning outcome for the students as a participant and a peer-reviewers. The results from the questionnaires were also discussed in the qualitative interviews. Since the questionnaire was anonymous, there was no opportunity to connect the questionnaires with the interviews. We have conducted qualitative interviews with 15% of the enrolled students for each course.

Challenges and opportunities for innovating assessment practices

At the beginning of developing the assessment tool, one of main challenge was establishing the criteria for assessment. The most complicated aspect was how to define their descriptive scope as well as the nuance of weighting. The criteria were adjusted during the process, some becoming more detailed and others being removed. The process and several iterations resulted in grouping the criteria in a more general group, which can be linked to the course taxonomy. Based on the work done in this paper, new courses can easily employ these criteria with minimal configuration. Working with the criteria has also given the lecturer better understanding in how to develop better assignments, i.e. assignments with a clear distinction of levels.

From the user/students'side, the responses regarding to the usability and stability of the software program are displayed in Figure 1 below. We notice a high number of positive responses regarding both variables, with more than 60% of the students being content or very content with the stability of the tool and about half of them with the usability. Considering this was the first iteration of the innovation, we consider these experiences as being satisfactory and good input for proceeding with further adjustments.

	Very Good	Good	Fair	Poor	Very Poor
Usability	*14,8 %*	*44,4 %*	*37,0 %*	*3,7 %*	*0,0 %*
Stability	*18,5 %*	*44,4 %*	*18,5 %*	*11,1 %*	*7,4 %*

Figure 3. Users response to the usability and stability of the assessment program.

Students' perceptions and experiences

The questionnaire data, from which we extracted answer to a number of core questions (Figure 4), to pin down students' experiences with the feedback provided through the innovation we implemented. The majority of the students were generally positive both to the way the feedback was given as well as the content. More that 90% of the students liked the form in which feedback was provided, namely, the way the feedback is structured and the way different elements are described. The responses indicate overwhelming positive engagement from the students with the feedback as a report or tool. They responded in the open comments field that this feedback helped them to understand their shortcomings regarding to what they are assessed for. An expectation was that students would use this information and deploy it to advance their knowledge and learning, but that appears not to be the case. We also mapped the students' prior expectations and how they felt the feedback is tailored for their assignment. For the extent of the feedback matching their own expectations on how they performed on the assignment the majority, 74%, responded good and very good. This could mean that the students are well informed about their own knowledge level, but not sure of what they are being measured, given their answers that they appreciate guidance about what is expected of them.

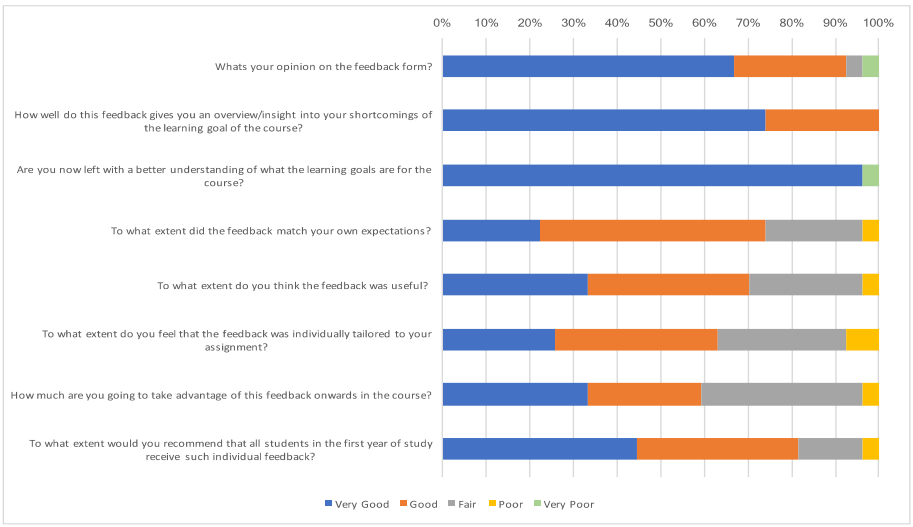

Figure 4. Overview of questionnaire items responses on expectations and perceptions

Based on the questionnaire answers, the students appear to have a good understanding of their own level of knowledge and how they performed on the assignment, self-efficacy, but lack the context of what and how they are to be assessed. Especially, it seems to be of great interest for the students how their assignment is assessed in term of / in context of the learning goals of the course. Finally, with regard to the tailoring of the feedback to their assignment and performance, almost 60% of the student rated it as very good and good, which indicated that the automated assessment program is able to provide feedback that is perceived as individualized and appropriate.

Connection to the main conference theme.

This paper presented a case study that involved the development and testing of a software program and criteria for formative individualized assessment of exams and automated feedback in computer science education. The innovative software collects data within several dimensions, by means of the hand-in assignments, the evaluators' response and the formative feedback. The automated data allows for employing an algorithms-based analysis, which provide the lecturer with input regarding new ways and tools to improve the learning gain of the students. Furthermore, the collected data and the logs from the software program can be used through machine learning and AI to teach the program able to self-determine the grade. This innovative assessment program gives a unique and efficient way to collect the much-needed dataset to potentially perform and verify new algorithms within Big Data and AI.

In addition, the students' perspectives to feedback on their exams and the experiences with the software program and the feedback provided were explored. The findings indicate positive experiences and students benefiting from the feedback, showing that providing the students with explicit feedback in addition to regular summative evaluation (grades and point sum) stimulated their understanding and awareness of their knowledge level and performance (see also Greenhow, 2015 and Siddiki et al., 2010). Besides, it triggered ideas and reflections related to their future learning steps and approaches, which follows the principles of feedforward. The program supports providing feedback specifically aimed at the students' professional skills, thus triggering focused alternatives for future learnings steps. In addition, the study also provides insights into how automated feedback generated through the use of criteria can be organized by means of a dedicated software program. The students' reports related on usability and the stability of the program indicate that the software served the envisioned purposes, which gives basis to concluding that these localized trials can be scaled up and the developed criteria can be further refined.

At the level of the educational practice, the method employed provides an innovative tool available for the teacher and the evaluator to provide the students with feedback. The developed system involves teacher's work to define in writing (phrase) what given values of criteria mean, with several criteria leading a matrix of possible feedback. The different feedback will make it easier to calibrate the evaluators across the disciplines. That way, not only the students benefit from this approach, but also the teachers/evaluators gain better understanding of what is required from the students in the various assignments and future learning. Furthermore,

the system provides possibility to perform reliable assessment, i.e., calculate degrees of agreement between evaluators, which can represent a considerable achievement.

References

Baker, D.J., Zuvela, D. (2013), Feedforward strategies in the first-year experience of online and distributed learning environments, Assessment & Evaluation in Higher Education, Vol. 38, No. 6, pp 687-697.

Bloom, B.S. (Ed.). Engelhart, M.D., Furst, E.J., Hill, W.H., Krathwohl, D.R. (1956). Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain. New York: David McKay Co Inc.

Conaghan, P. and Lockey, A., (2009). Feedback to feedforward, *Notfall+ Rettungsmedizin*, Vol. 12, No. 2, pp.45-48.

Ellery, Karen (2007). Assessment for learning: a case study using feedback effectively in an essay-style test, *Assessment and Evaluation in Higher Education*, 33(4), 421-429.

Greenhow, M (2015). Effective computer-aided assessment of mathematics; principles, practice and results. *Teaching Mathematics and its Applications: An International Journal of the IMA,* 34(3), 117 - 137.

Higgins, R., Hartley, P., Skelton, A. (2001), Getting the Message Across: The problem of communicating assessment feedback, *Teaching in Higher Education*, Vol. 6, No. 2, pp. 269-274.

Lewis, D. J. A. &. Sewell, R. D. E (2007). Providing Formative Feedback From a Summative Computer-aided Assessment. *American Journal of Pharmaceutical Education*, 71(2), https://doi.org/10.5688/aj710233

Malmi, L., Korhonen, A. (2004), Automatic Feedback and Resubmissions as Learning Aid, IEEE International Conference on Advanced Learning Technologies, pp. 186-190.

Mirmotahari, O. & Berg, Y. (2017), *Individuell «automagisk» tilbakemelding på skriftlig eksamen.* Nordic Journal of STEM Education, Vol. 1, No. 1 pp. 287-293

Moesby, E. (2002), From Pupil to Student – a Challenge for Universities: an Example of a PBL Study Programme, Global Journal of Engineering Education, Vol. 6, No. 2, pp. 145-152.

Sadler, D.R. (2010), Beyond feedback: developing student capability in complex appraisal, Assessment & Evaluation in Higher Education, Vol. 35, No. 5, pp 535-550.

Hunter-Barnett, S., Murrin-Bailey, S., Should Audio Feedback Be Used Because It Is Easily Available or For Reasons of Pedagogy? In Proceedings of the 2nd International Conference on Computer Supported Education, pp 60-64.

Jiménez-González, D., Álvarez, C., López, D., Parcerisa, J.M., Alonso, J., Pérez, C., Tous, R., Barlet, P., Fernández, M., Tubella, T. (2008), Work in Progress–Improving Feedback Using an Automatic Assessment Tool, IEEE Frontiers in Education Conference, Session S3B.

Roderick, M., Holsapple, M., Kelley-Kemple, T., Johnson, D.W. (2014), From High School to the Future: Getting to College Readiness and College Graduation, Society for Research on Educational Effectiveness.

Siddiqi, R., Harrison, C.J., Siddiqi, R., Improving Teaching and Learning through Automated Short-Answer Marking, IEEE Transactions on Learning Technologies, Vol. 3, No. 3, pp 237-249.

van der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. Computers & Education, 58(1), 263-272.

van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis. Review of Educational Research.