Evaluating students' self-assessment in large classes

Jokke Häsä[1], Johanna Rämö[2] and Viivi Virtanen[3]

University of Helsinki, Finland; [1]jokke.hasa@helsinki.fi, [2]johanna.ramo@helsinki.fi, [3]viivi.virtanen@helsinki.fi

This study is part of an ongoing larger project concerning student self-assessment skills in university courses. We have developed a method enabling large cohorts of students to assess their own learning outcomes and to give their own course grades with the help of an automatic verification system. This paper explores the question of accuracy, namely, whether the self-assessed grades correspond to the students' actual skills, and how well the automatic system can pick up issues in the self-assessment. Based on an expert's evaluation of the skills of two students, we conclude that although for large part the model works as intended, there are some cases where neither the self-assessment nor the computer verification seem to be accurate.

Keywords: Self-assessment, Assessment for learning, Digital assessment, Accuracy, Large classes

Introduction

The ability to judge the quality of one's own work is one of the core skills that should be developed during university studies. Self-assessment has been viewed as a valuable assessment process through which student can learn to understand the expectations, criteria and standards used in assessment, and further, to be able to regulate own learning and acquire skills for lifelong learning (Falchikov & Boud, 1989; Kearney, Perkins, & Kennedy-Clark, 2016). However, academic community seem to be resistant to change the prevailing assessment practices focusing on testing and grading, and practices such as self-assessment are scarcely implemented at course level (Boud et al., 2018; Postareff, Virtanen, Katajavuori, & Lindblom-Ylänne, 2012).

In this paper, we draw attention to assessment practices in university first-year mathematics by examining an implementation of student self-assessment processes into large class setting. During this process, students frequently evaluated the quality of their learning outcomes, received feedback on their performance, and finally decided their own grades according to particular criteria. The intended learning outcomes were made transparent through a rubric including both content knowledge and generic skills, such as writing mathematics. The digital environment gave opportunity for monitoring learning process and giving real-time formative feedback in line with previous research on online assessment (Ćukušić, Garača, & Jadrić, 2014; Gigandi, Morrow, & Davis, 2011; Ibabe & Jauregizar, 2010), and further, it formed a basis of assessment of the student's progress. The emphasis of self-assessment was in developing student capability in making evaluative judgements (Ajjawi, Tai, Dawson, & Boud, 2018) and building their metacognition skills (Mok, Lung, Cheng, Cheung, & Ng, 2006), so that the students' ability to self-regulate their learning for current and future learning would improve. We fill the gap in research by showing how, in the case of summative self-assessment, the problems aroused by large class setting were resolved by using digital and automatic verification and real-time feedback.

Self-assessment as a Tool for Learning

Self-assessment can be defined as a process during which student evaluate their own achievements and judge about their own performance (Falchikov & Boud, 1989). The judgements students make are based on information and evidence about their own performance collected from various sources (Yan & Brown, 2017). In this paper, we refer to self-assessment as a process in which the students evaluate their own progress and performance and give justifications for the result of their evaluation according to teacher-given criteria showing intended learning outcomes.

The use of self-assessment has been shown to improve student engagement and motivation (e.g. Andrade & Du, 2007; Mok et. al, 2006), self-efficacy (Kissling & O'Donnell, 2015) and academic performance (Ibabe & Jauregizar, 2010), while the ability to self-assess is reportedly intermingled with ability to self-regulate own learning (Panadero, Brown, & Strijbos, 2016) and with life-long learning skills (Kearney et al., 2016). Consequently, the literature encourages the use of self-assessment for formative purposes. Research shows that in large class settings digital environments with effective formative online assessment can foster a learner-centred focus and engagement in learning (Gigandi et al. 2011; Ibabe & Jauregizar, 2010). Recent results show that online self-assessment can also improve students' academic success (Ćukušić et al. 2014). However, the debate concerning students generating their own grades by self-assessing their own work is more complicated and constantly questioned (Boud et al., 2018; Tejeiro et al., 2010). One of the main challenges regarding self-assessment for grading is the question of accuracy: How can we be sure that students' grades are valid and reliable?

The Question of Accuracy

Many studies have found high correlations between self- and teacher-ratings (Falchikov & Boud, 1989; Kearney et al., 2016). The results indicate that students are able to make reasonable accurate judgements if they are properly provided with training and background information to the process. Also, students vary in their capability to evaluate their performance e.g., high achievers tend to underestimate their performance whereas low achievers tend to overestimate it (Boud & Falchikov, 1989; Boud et al., 2013; Kearney et al., 2016). However, the accuracy of student self-assessments can be improved through using criteria and standards (Andrade & Du, 2007), while students need to have multiple opportunities for practising self-assessment in relation to given criteria, with feedback to help calibrate the judgements (Hosein and Harle, 2018; Kearney et al., 2016). On the other hand, Boud, Lawson and Thompson (2013) argue that increase in accurate self-assessment is not immediately transferable, because standards and criteria are somewhat domain-specific. Hence, we suggest that in order to understand the expectations, criteria, and disciplinary standards of mathematics, and to develop capabilities to make accurate and realistic assessments on own learning processes and outcomes, it is required that self-assessment processes are implemented in first-year university mathematics. However, in large class setting, typical to that learning context, the challenge how to give evidence-based feedback for improving the accuracy needs to be resolved.

The DISA model

This study is part of a research project centred around an assessment model called DISA (**Di**gital **S**elf-**a**ssessment). In the model, students assess their own learning outcomes throughout the course by using a detailed rubric articulating the subgoals of the ultimate intended learning outcomes. Learning goals and criteria are clearly identified, and through self-assessment activities the students are actively engaged with them. Evidence of learning is elicited during the course, and students receive feedback for their self-assessment from an automatic digital system.

The feedback is generated in the following way. Every course task has been linked with the learning objectives it is supporting. This enables the automatic system to compute, based on the student's coursework, an index from 0 to 1 for each learning objective. This index estimates how well the student has acquired the learning objective. From these indices, the system then computes tentative grades in each course topic. These tentative grades are compared to the student's self-assessed grades, and the student is advised either to consider a higher or a lower grade for themselves.

In addition, self-assessment is used for summative purpose in the end of the course, as the students self-assess and justify how well they have achieved the intended learning outcomes, and proceed in deciding their own course grade based on the self-assessment. In order to prevent abuse of the self-assessment process, the system described above is used to verify the validity of the course grades. If the self-assessed topic grades differ too much from the computed ones, the student's final course grade is disputed. Earlier results imply that the model supports students in using deep learning approach, and encourages them to study for themselves, not for an exam (Nieminen, Rämö, Häsä, & Tuohilampi, 2017).

Aim of the Study

This study aims at gaining a better understanding of the use of self-assessment as an integral part of assessment in a large first-year university mathematics course. In the course context, self-assessment is used to give students an opportunity to think metacognitively about their learning. We hypothesise that student active engagement into self-assessment processes is enhanced if these processes are valued in grading, but then, the question of accuracy needs to be resolved. This question is two-fold: firstly, we are interested in the validity of the student grades, in other words, whether they reflect true learning, both in content knowledge and domain-specific generic skills such as writing mathematics. Secondly, we need to examine the reliability of the automatic verification system: can it spot the cases where self-assessment is inaccurate? The research questions in this study are:

1. How do the students' evaluations of their own skills compare with evaluations performed by the automatic verification system?
2. How does an expert judge the student's acquired skills in cases where the automatic verification disagrees with student's self-assessment?

Method

This study uses data collected from students taking a first year mathematics course at a major research-intensive university in Finland. The second author was the lecturer for the course. The course was a proof-based linear algebra course dealing with finite-dimensional vector spaces, and it lasted for seven weeks (half a term). During the course, students were given weekly problems to solve, part of which were assessed and given feedback on. Some of the tasks were assessed by the tutors, some by an automatic assessment system called Stack (Sangwin, 2013). Some tasks were also peer-reviewed.

The course was not graded with a final exam, but grading was done by self-assessment using the DISA model. The self-assessment was based on a detailed learning objectives matrix prepared by the teacher. The learning objectives were divided into 10 topics: six content-specific and four generic skills topics, and the students were asked to give themselves a grade from 0–5 in each of these topics, 0 meaning fail. They were also asked to write down reasons for choosing that grade. In the end of the course, students chose their own final grade. They were left to decide by themselves how to combine the grades from the different topics. The DISA system was used to verify the final self-assessment.

It is worth noting that, in the Finnish context, although the teacher is responsible for the course grades, these can be awarded by any means the teacher chooses. There is little fear of distorting the grades, as the final grade of a first-year mathematics course carries very little weight in the final outcome of a student's study programme. Also, all courses and exams can be usually retaken as many times as the student wishes.

The participants of this study were 158 students who took the linear algebra course described above, gave themselves their own grades using the DISA model, and gave consent for using their data. Most of the students were majoring in either mathematics, mathematics education or some other field related to mathematics such as computer science, physics or chemistry. Most students were first year students, but the cohort included also older participants, up to post-doctoral level.

We narrow our study to two of the ten learning objective topics of the course: (1) "Matrices" (content-specific) and (2) "Reading and writing mathematics" (generic skill). These two topics were chosen since both are among the most central topics of the course and there were relatively many tasks linked to them. Also, we wanted to compare self-assessments on a content-specific topic with those on a generic skill. Henceforth, these topics are abbreviated as [M] and [RW]. Examples of learning objectives pertaining to these topics are given in Table 1.

Topic	Skills corresponding to grades 1-2	Skills corresponding to grades 3-4	Skills corresponding to grade 5
Matrices [M]	I can perform basic matrix operations and know what zero and identity matrices are	I can check, using the definition of an inverse, whether two given matrices are each other's inverses	I can apply matrix multiplication and properties of matrices in modelling practical problems
Reading and writing [RW]	I use course's notation in my answers	I write complete, intelligible sentences that are readable to others	I can write proofs for claims that concern abstract or general objects

Table 1: Part of the learning objectives matrix of the course. In total, there were 10 topics and 10–15 learning objectives in each topic

To answer Research question 1, we compared the grades students gave themselves on the two topics in the final self-assessment with the results of the automatic verification of that self-assessment. The computations were done with R version 3.5.0. For Research question 2, coursework and final self-assessment of two students whose self-assessment was poorly in line with the automatic verification were chosen for closer inspection. In this manuscript, we call them Student A and Student B. The two students' anonymised answers to all of the written tasks as well as their Stack exercise points were analysed by the second author. This author was also the teacher of the course and can be regarded as an expert in the subject. When the expert was grading the students, she did not know how the students had assessed themselves. The expert read every written solution the student had submitted, and evaluated which learning objects in topics [M] and [RW] the student had reached.

Every time the expert could see the student mastering a learning object, she made a note in the learning objectives matrix. After that, there were learning objectives for mastering of which the student had not provided any evidence in the written solutions. The expert then looked at the Stack exercises that were linked to these learning objectives to see how many points the student had received from them. She used the information in determining whether the student had reached the remaining learning objectives. When the expert had considered each learning objective, she awarded the student a grade in both topics by looking from the learning objectives matrix which grade the reached learning objectives corresponded to. In borderline cases, the expert used her expertise as a mathematician and teacher of the course. For the topic "Reading and writing mathematics", the expert could only evaluate the student's skills in writing as there were no tasks that were linked to reading skills.

Results

Research question 1: comparison of self-assessed grades with automatic verification

The distributions of the self-assessed grades in the two topics [M] (Matrices) and [RW] (Reading and writing mathematics) are shown in Table 2. We see that the students gave the grades 3 and 4 more often for [RW] than for [M], but the top grade 5 was more common in [M] than in [RW].

Grade	1	2	3	4	5
[M]	3	10	25	47	73
[RW]	2	10	37	58	51

Table 2: Frequencies of each grade in the two topics.

The computer verification system computed tentative grades for the two topics for each student. The distribution of differences between the computed grade and student self-assessed grade are reported in Table 3.

Difference	-5	-4	-3	-2	-1	0	1	2
[M]	1	1	5	20	26	86	19	0
[RW]	0	2	6	15	20	75	34	6

Table 3: Frequencies of the differences: computed grade minus self-assessed grade in the two topics.

In [M], there are 86 matches, 53 cases in which the self-assessed grade was higher than the computed grade (negative difference), and only 19 cases in which the self-assessed grade was lower (positive difference). In [RW], there are 75 matches, 43 cases in which the self-assessed grade was higher, and

40 cases in which the self-assessed grade was lower. In both topics, between 81-83 % of self-assessed grades lie within 1 grade point from the computed grade.

Research question 2: Expert opinion in conflicted cases

Student A's self-assessed grades were lower than the computed ones. For both topics, the self-assessed grade was 4 and computed grade 5. The expert's evaluation agreed with the computed grades. The expert observed that Student A had done almost all tasks during the course. Even though not all the answers were correct, all the learning objectives in topic [M] were fulfilled. Students were asked to make corrections to some tasks, and student A had always resubmitted solutions written in good mathematical style. The student's explanations were concise and readable, and the student was able to construct proofs concerning abstract mathematical objects. Based on this, the expert's grade for topic [RW] was 5.

Student B's self-assessed grades were greater than the computed ones. For topic [M], the self-assessed grade was 5 and the computed grade 3. The expert's evaluation yielded grade 4, that is, something in between. For [RW], the self-assessed grade was 3 and the computed grade 1. The expert's evaluation agreed with the self-assessed one. The expert observed that Student B had submitted only a fraction of the course tasks. However, the expert was able to evaluate from the solutions that Student B accomplished almost all learning objectives in [M]. Some of Student B's skills were shown in the intermediate steps of tasks that were not directly linked to topic [M]. For example, the student determined whether given vectors are linearly independent by forming a system of linear equations and calculating the determinant of the coefficient matrix. This showed that the student knew how invertibility of matrices is connected to the number of solutions of a system of linear equations even though the topic of the task was linear independence. Student B had not corrected any solutions when encouraged to. According to the expert, the student reached partially all the learning objectives in [RW], but did not fully master any of them, not even the ones corresponding to grade 1. For example, the student mixed up equivalence arrows with equality signs, wrote long, confused sentences and used "if–then" structures inside a proof in the place of assumptions and conclusions. However, the overall structures of the proofs were correct. Based on this, the expert's interpretation was that the student's grade for [RW] was 3.

Discussion

In this study, a new model of determining course grades via self-assessment was examined with a focus on the accuracy of the self-assessed grades. The students gave themselves grades in all course topics, and these grades were automatically verified by comparing them against the course work the students had done. We analysed the results of the verified self-assessment in two topics, one content-specific topic (matrices) and one subject-related generic skill (reading and writing mathematics).

The students' self-assessment agreed well with the automatic verification. Most discrepancies are within one grade point, which can be explained by the coarseness of the grading scale: the "real" skill level is often between two grade points and must be forced to one or the other direction. This is true for any assessor, be it student, computer or teacher. The high agreement is not surprising, as previous studies have shown that explicit criteria and standards support self-assessment, as does frequent practice and feedback (Andrade & Du, 2007; Kearney et al., 2016). It remains to be studied how great an effect the feedback that the students received for their self-assessment exercises had on their final self-assessment.

The students gave fairly good grades to themselves in both examined topics. For reading and writing mathematics, the grades were more concentrated around the second-best grade, whereas for matrices, the top grade was clearly the most common grade. Perhaps it was easier for the students to understand the learning objectives as well as recognise their achievements in the mathematical topic, and without clear evidence for mastery, they were hesitant to award themselves the best grade in a generic skill. Our results could be understood in the view of previous results (Falchikov & Boud, 1989) showing that in science courses, self-assessment was more accurate that in other fields.

We examined more closely two students whose self-examined and computed grades differed. In the first case, self-evaluated grades were below the computed ones. The expert's evaluation agreed with the computed grade. The student was a high achiever, and from previous studies we know that such students tend to underestimate their performance (Boud et al., 2013; Kearney et al., 2016). In the second case, the self-evaluated grades were above the computed ones. The expert's evaluation was between the two for the mathematical topic and agreed with the self-assessed grade for the generic skill. In this case, the student had skipped many tasks which made it difficult for the automatic system to estimate the grade fairly. Also, the expert noted that the student seemed to have some skills from all grade categories in the learning objectives matrix, but not to have fully reached any. This kind of case would be very difficult for the automatic verification system to estimate correctly.

The study used a method in which an expert evaluated students' skills based on all the work they had done on the course, evaluating against the intended learning outcomes, not by grading individual tasks. The method suffered from some of the maladies related to teacher evaluation, such as time restriction and personal bias. The accuracy of teacher-grading is not an issue to be taken as obvious truth (Brown, 1997). One should also note that neither the expert nor the automatic system were able to evaluate students' reading skills even though they were included in the self-assessed grades.

This study opened a new way to critically examine a self-assessment model as a viable option for grading students. We did not find any fundamental problem with reliability. However, at least in one of the studied cases, the verification system did not estimate the student's skills very well. A larger sample needs to be studied in order to find out whether such issues are common. Also, we need to study students' written justifications for their grades in order to better understand what is involved when the self-assessment process does not go as intended.

References

Ajjawi, R. Tai, J., Dawson, P. & Boud, D. (2018). Conceptualising evaluative judgement for sustainable assessment in higher education. In *Developing Evaluative Judgement in Higher Education* (pp. 23–33). Routledge.

Andrade, H. & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment & Evaluation in Higher Education*, *32*(2), 159–181.

Boud, D., Lawson, R., & Thompson, D. (2013). Does student engagement in self-assessment calibrate their judgement over time?. *Assessment & Evaluation in Higher Education*, *38*(8), 941–956.

Boud, D., Dawson, P., Bearman, M., Bennett, S., Joughin, G., & E. Molloy. (2018). Reframing assessment research: through a practice perspective. Studies in Higher Education, *43*(7), 1107–1118.

Brown, G., Bull, J., & Pendleburry, M. (1997). *Assessing student learning in higher education.* London: Routledge.

Ćukušić, M., Garača, Ž., & Jadrić, M. (2014). Online self-assessment and students' success in higher education institutions. Computers & Education, *72*, 100–109.

Falchikov, N. & Boud, D. (1989) Student Self-Assessment in Higher Education: A Meta-Analysis. *Review of Educational Research, 59*(4), 395–430.

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: a review of the literature. Computers & Education, *57*(4), 2333–2351.

Hosein, A. & Harle, J. (2018). The relationship between students' prior mathematical attainment, knowledge and confidence on their self-assessment accuracy. *Studies in Educational Evaluation, 56*, 32–41.

Ibabe, I. & Jauregizar, J. (2010). Online self-assessment with feedback and metacognitive knowledge. *Higher Education, 59*, 243–258.

Kearney, S., Perkins, T. & S. Kennedy-Clark (2016). Using self-and peer-assessment for summative purposes: analysing the relative validity of ASSL (Authentic Assessment for Sustainable Learning) model. *Assessment & Evaluation in Higher Education, 41*(6), 843–861.

Kissling. E. & O'Donnell, M. (2015). Increasing language awareness and self-efficacy of FL students using self-assessment and the ACTFL proficiency guidelines. *Language Awareness*, *24*(4), 283–302.

Mok, M.M.C., Lung, C.L., Cheng, D.P.W., Cheung, R.H.P. & Ng, M. L. (2006). Self-assessment in higher education: experience in using a metacognitive approach in five case studies. *Assessment & Evaluation in Higher Education, 31*(4), 415–433.

Nieminen, J. H., Häsä, J., Rämö, J., & Tuohilampi, L. (2017). Replacing Exam with Self-Assessment: Reflection-Centred Learning Environment as a Tool to Promote Deep Learning. Proceedings of the 20th Meeting of the MAA Special Interest Group on Research in Undergraduate Mathematics Education. San Diego: RUME

Panadero, E., Brown, G.T.L. & J.W. Strijbos. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational Psychology, 28*(4) 803–830.

Postareff, L., Virtanen, V., Katajavuori, N., & Lindblom-Ylänne, S. (2012). Academics' conceptions of assessment and their assessment practices. *Studies in Educational Evaluation, 38*(3–4), 84–92.

Sangwin, C. (2013). *Computer aided assessment of mathematics*. OUP Oxford.

Yan, Z. & Brown, G. (2017). A cyclical self-assessment process: towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education, 42*(8), 1247–1262.