Evaluating students' self-assessment in large classes

DRAFT

Häsä, J., Rämö, J. & Virtanen V.*
University of Helsinki
Aalto-University*

Abstract

This study is part of an ongoing larger project concerning student self-assessment skills in university courses. In particular, we have developed a method enabling large cohorts of students to assess their own learning outcomes and to give their own course grades. This paper explores the question of accuracy, namely, validity and reliability of the self-assessments created with the method in one particular mathematics undergraduate course.

Introduction

Research in assessment shares a consensus that the ability to judge the quality of one's own work is one of the core skills that should be developed during university studies. In regard to this, 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; Crisp 2012; Kearney et al. 2016; Panadero et al. 2016). However, academic community seem to be highly 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 et. 2012; Halinen et al. 2015). Meanwhile, the higher education should provide the graduates with generic skills that are globally viewed as crucial and often transferable competences for future employment (Jääskeläinen et al. 2018). But, many of the generic skills are largely bound to their contexts and fields, and in consequence, we suggest that it need to be explored if the kind of skills are accurately assessed in higher education mathematics.

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 the self-assessment processes in the course, 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 explicitly made transparent through a rubrics included both content knowledge and domain-specific generic skills, such as writing mathematics. The emphasis of self-assessment was in developing student capability in making evaluative judgements (Ajjawi et al. 2018) and in building the metacognition of the learners (Mok et al. 2006), so that the students' abilities to self-regulate their learning for current and future learning will be improved through systematic self-assessment. Here, we fill the gap in research by showing how the problems aroused by large class setting were resolved by using digital and automatic verification and feedback. In this paper, we especially focus on exploring the accuracy of students' self-assessments.

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 (Boud and Falchikov, 1989). Recently, self-assessment has been defined as follows: "self-assessment is the qualitative assessment of the learning process, and its final product, realized on the basis of pre-established criteria'' (see Panadero et. al. 2016). In accordance, 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 (Andrade & Du 2007; Ćukušić et al. 2014; Mok et. al 2014, Nikou & Economides 2014), 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 et al. 2016) and with life-long learning skills (Boud 2000; Crisp, 2012; Kearney et al. 2016). Consequently, the literature encourages the use of self-assessment for formative purposes, but 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 and Boud 1989; Asikainen et al. 2014; 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. Results also indicate that students vary in their capability to evaluate their performance e.g., high achievers tend to underestimate their performance whereas low achievers tend to overestimate their performance (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 practicing self-assessment in relation to given criteria, with feedback to help calibrate the judgements ((Hosein and Harle, 2018, Kearney et al. 2016; Yusel et al. 2017). Findings (Boud et al. 2013) suggest that students become more effective in criteria-based judgements both within a subject and a range of subjects over time. However, they argued that increase in accurate self-assessment is not immediately transferable, because standards and criteria are somewhat domain-specific and understanding of the expect assignment need to be gained whenever confronted to a new subject matter. Hence, we suggest that in order to understand the expectations, criteria, and disciplinary standards of the 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, 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 centered around an assessment model called DISA (DIgital Self-Assessment). In the model, students assess their own learning outcomes throughout the course by using a detailed rubrics articulating the subgoals of the ultimate intended learning outcomes of the course. 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. Self-assessment is not focused on individual course tasks but students make evaluative judgements directly about their progress in reaching the intended learning outcomes.

In addition, self-assessment is used for summative purpose and in the end of the course, as the students self-assess and justify the how well they have achieved the intended learning outcomes, and proceed in deciding their own course grades based on the self-assessment. In order to prevent abuse of the self-assessment process, the automatic digital system is used to verify the validity of the grades. In the system, every course task has been linked with the learning objectives it is thought to support. This enables the system to compute a tentative grade for each topic from the student's assessed and non-assessed coursework. The tentative grades are compared with the student's self-assessed grades, and if there are too large discrepancies in one direction or the other (or both), the student's final grade is disputed. If the student's written justifications were not enough to justify the discrepancy, the student was asked to take a short automatically assessed test or suggest themselves another grade. Otherwise the student's self-assessed grade was awarded to the student.

It is worth noting that, in the Finnish context, although the teacher is finally responsible for the course grades, these can be awarded by any means the teacher chooses. Hence, it is possible to hand a lot of power over to the students. There is also little fear of distorting the grades by doing this, as the final grade on a first-year mathematics course carries very little weight in the final outcome of a student's study programme. This is because there are several other courses assessed in diverse ways, and all courses and exams can be retaken as many times as the student wishes.

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 class. In the course context, self-assessment is used to give students an opportunity to think metacognitively about their learning. However, 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, namely, 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

Context

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).

Teaching in the course was student-centred and based on the Extreme Apprenticeship Model (Vihavainen, Paksula, & Luukkainen, 2011; Rämö, Oinonen & Vihavainen, 2016). In this model, students take part in activities resembling those of experts. New topics are introduced by scaffolded tasks, and students start working on them outside lectures. Students are offered guidance in an open, drop-in learning space by peer tutors, who receive training throughout the course.

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 (Nicol, 2014). In the tasks assessed by the tutors, special attention was paid to readability and good mathematical style, and students were encouraged to rewrite their solution for a second assessment if it did not meet the goals or was incorrect.

The course was not graded with a traditional final exam, but assessment 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 1–5 in each of these topics. They were also asked to write down reasons for choosing that grade. In the end of the course, students chose their own final grades. They were left to decide by themselves how to combine the grades from the different topics. The automatic DISA system was used to verify the final self-assessment.

Participants

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. However, there were also students of other disciplines, such as social sciences. Most students were first year students, but the cohort included also older participants, up to post-doctoral level.

Data Collection and Analysis

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 ??.

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

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	In my solutions, I write complete, intelligible sentences that are readable to others	I can write proofs for claims that concern abstract or general objects

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. We used Kendall's tau-b as a non-parametric correlation measure (Kendall, 1938) when comparing grades, as the distances between adjacent grades are not necessarily identical.

For Research question 2, coursework and final self-assessment of two students were chosen for closer inspection to explore reasons that might explain a poor result in the automatic verification. In this manuscript, we call them Student A and Student B. For Student A, automatic verification suggested higher grades in both topics, and for Student B lower.

The two students' anonymised answers to all of the written tasks of the course were analysed and their points awarded to each Stack exercise examined by the second author. This author was also the teacher of the course and can be regarded as an expert in the subject. 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 those exercises, and used that information in evaluating 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.

When the expert was grading the students, she did not know how the students had assessed themselves. Regarding the topic "Reading and writing mathematics", as there were no tasks that were linked to reading mathematics, we could only evaluate students' skills in writing.

Results

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

The distributions of the self-assessed grades in the two topics "Matrices" (henceforth abbreviated [M]) and "Reading and writing mathematics" (henceforth abbreviated [RW]) are shown in Table ???. 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 ??: 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 ??

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

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

We see that the computer and student grades agree well. 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.

To quantify the agreement between computed and student grades, Kendall’s tau coefficient was computed between the two variables for both topics. We obtain $\tau = 0.61$ for [M] and $\tau = 0.57$ for [RW]. Both are relatively high values, but we see that agreement was greater for [M].

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. Student A had made corrections to the solutions when asked to, and the resubmitted solutions were 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 linked to the number of solutions of a system of linear equations even though the topic of the task was linear independence. Also, Student B had solved some of the problems using techniques not presented in this course.

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 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 students grade for [RW] is 3.

The results of the case study are summarised in Table ??.

	Student A			Student B		
	Self-assessed	Computed	Expert	Self-assessed	Computed	Expert

[M]	4	5	5	5	3	4
[RW]	4	5	5	3	1	3

Table ??: Self-assessed grades, grades computed by the automatic verification system, and grades obtained from expert evaluation for the two case subjects in the two topics.

Discussion

All in all, the students' self-assessment agreed very 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. 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. Probably, it is more easy for students to make accurate assessment on own learning if the criteria and intended learning outcomes are unambiguous. It might be that learning objectives and criteria for reading and writing may have been slightly difficult both to describe (for the teacher) and to understand (for the students).

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 students was a high achiever, and from previous studies we know that such students tend to underestimate their performance (Boud & Falchikov 1989; Boud et al. 2013; Kearney et. al 2016). The precise reasons for this underestimation is left for further study. However, a preliminary look at the reflections the student wrote in their self-assessment suggests that the student had felt that one needs to reach all the learning objectives perfectly in order to deserve the best grade.

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 as there were not much data. 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.

Limitations and Further Study

The study used a method where 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 well-known maladies related to teacher evaluation, such as time restriction and personal bias. The expert noticed that after a few days break and after discussions with another author of the paper, she would perhaps think differently about the students' skills. Thus, the accuracy of teacher-grading is not an issue to be taken as a obvious truth (Brown, 1997). Bearing this in mind, in the future there should perhaps be more than one person to evaluate the students, and more than one iterations .

Other problems arose from the learning objectives. Some of the ones concerning reading and writing were rather ambiguous. Also, in some cases the learning objectives the student had reached were scattered across the matrix, and determining the grade was difficult for the expert. One could imagine that the students face similar problems when assessing themselves. 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 self-assessed grades. This can cause discrepancies between self-assessed grades and both computed grades and expert's evaluation.

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, at least in the low-stakes environment. 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, and on the other hand, whether the automatic system will let students overestimate their grades so much that the grades could become discredited. 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

Vihavainen, Paksula, & Luukkainen, 2011; Sangwin, 2013
Rämö, Oinonen & Vihavainen, 2016).

Nicol, David; Thomson, Avril; Breslin, Caroline. 2014. Rethinking feedback practices in higher education: a peer review perspective. =? : Assessment & Evaluation in Higher Education. Jan2014, Vol. 39 Issue 1, p102-122. 21p. 1 Chart. DOI:.

Ajjawi, R. Tai, J., Dawson, P. & Boud, D. (2018). Conceptualising evaluative judgement for sustainable assessment in higher education. In: Boud et al. (edit.). *Developing Evaluative Judgement in Higher Education*. Oxford, UK. Routledge. pp. 202 pp.

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

Asikainen, H., Postareff, L., Heino, P. & Virtanen, V. (2014). Peer assessment in a large introductory class of gene technology. Studies in Continuing Education, 43, 197–205.

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22*(2), 151-167.

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.

Crisp, G. T. (2012). Integrative assessment: Reframing assessment practice for current and future learning. *Assessment & Evaluation in Higher Education, 37*(1), 33-43.

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.

Halinen, K., Ruohoniemi, M., Katajavuori, N., & Virtanen, V. (2013). Life science teachers' discourse on assessment:A valuable insight into the variable conceptions of assessment in higher education. *Journal of Biological Education,* 42, 16-22.

Hosein, A. & Harle, J. (2018). The relationship between students' prior mathematical attainment, knowledge and confidence on their self-assessment accuracy. *Studies in Educational Evaluation,*5 , 32-41.

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

Jääskelä, P., Nykänen, S. & Tynjälä, P. Models for the development of generic skills in Finnish higher education. *Journal of Further & Higher Education*, 42 (1)130-142.

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. & M.L.Ng. (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.

Nicol, D., Thomson, B. & Breslin, B. (2014) Rethinking feedback practices in higher education: a peer review perspective. Assessment & Evaluation in Higher Education, 39(1)102-122.

Nikou, S. A. & Economides, A. A. (2016). The impact of paper-based, computer-based and mobile-based self-assessment on students' science motivation and achievement. *Computers in Human Behavior,* 55, 1241–1248.

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.

Rämö, Oinonen & Vihavainen, 2016

Tejeiro, R. A., Gómez-Vallecillo, J. L., Romero, A. F., Pelegrina, M., Wallace, A., & Emberley, E. (2012). Summative self-assessment in higher education: Implications of its counting towards the final mark. Electronic Journal of Research in Educational Psychology, 10(2), 789-812.

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.

Yucel.R., Bird, F. L., Young, J., & Blanksby, T. (2014). The road to self-assessment: exemplar marking before peer review develops first-year students' capacity to judge the quality of a scientific report. *Assessment & Evaluation in Higher Education,* 39, 8, 971-986.