Accuracy of self-assessments in higher education[1]

Joerg M. Haake[2], Niels Seidel[2], Marc Buchart[2], Heike Karolyi[2],
Regina Kasakowskij[i][2]

Abstract Self-assessment serves to improve learning through timely feedback on one's solution and iterative refinement as a way to incrementally improve one's competence. The SelfAssess-plugin developed in Moodle provides unlimited opportunities for students to create and assess solutions to self-assessment questions. A field study examined how 131 students used voluntary self-assessment questions in an online course in B.Sc. Computer Science, how accurate they were able to self-assess their solutions on instructor-defined criteria, and which question-related characteristics influence the ability of self-assessment. Results show the potential for providing scalable learning support. Fine-grained assessment criteria and freedom of solution input are recommended; a limited complexity of the expected solution is still a challenge.

Keywords: self-assessment accuracy, computer-assisted feedback, criteria-based self-assessment

1 Introduction

Self-assessments serve to improve learning through timely feedback on one's solution and iterative refinement as a way to incrementally improve one's competence. In 2020, we began to work on scalable self-assessments with high information feedback in a computer science (CS) course [Haa20] and continued this research through another field study. In winter term 2020/21, 131 students of a 3rd semester bachelor CS study program used self-assessment questions providing a formative high information feedback for long answer questions [Haa20], that will be further explored in this article. The self-feedback is based on feedback generated from their self-assessment according to the fulfillment of predefined criteria for their given answer.

Students can only make use of formative assessment, if their self-assessment is accurate enough and if they make use of feedback for improving their answer [HBZ13, BW09, WZH20]. In order to evaluate whether such self-assessment is practically useful, we examined the usage of voluntary self-assessment questions, the accuracy of student self-assessment, and the impact of task characteristics on self-assessment accuracy.

[1] This research was supported by the Research Cluster Digitalization, Diversity and Lifelong Learning – Consequences for Higher Education (D^2L^2) of the FernUniversität in Hagen.

[2] FernUniversität in Hagen, Research Cluster D^2L^2, Universitätsstr. 11, 58084 Hagen, Germany, {joerg.haake https://orcid.org/0000-0001-9720-3100, niels.seidel https://orcid.org/0000-0003-1209-5038, marc.buchart https://orcid.org/0000-0001-5668-7137, heike.karolyi https://orcid.org/0000-0003-2368-9851, regina.kasakowskij https://orcid.org/0000-0002-8587-9530}@fernuni-hagen.de

2 Related work

To support students with a scalable immediate feedback and repeatability, a new question type named self-assessment with high information feedback was presented as an approach in an online learning environment [Haa20]. Self-assessments in online learning environments offer learners a variety of exercise options to deal with learning goals, requirements and assessment criteria and therefore giving an opportunity to make use of a self-feedback to increase one's learning success [Har17]. Self-assessments are especially helpful, if students are also able to accurately assess their actual performance or at least improve their assessment after a repetition.

Many students tend to evaluate themselves inaccurately, in particular when they are at the beginning of their studies. Students with poorer results tend to overestimate their own performance [Lew10, TG+12, Car20, BLT15]. [Lew10] found that academically more competent students have a more accurate self-assessment than less competent students and that the self-assessment ability of a student does not improve over the duration of a semester. In contrast, [BLT15] found an improvement over a longer period of time. Giving students criteria to assess themselves and also providing a teacher's feedback clearly helped students to understand the criteria and standards in order to get a realistic perception of their performance. They also stated that disruptive patterns of assessment within a sequence of subjects can reduce convergence between student and tutor evaluation. In addition, [Car20] showed that students who initially had a high level of accuracy in their self-assessment worsened in their accuracy in a subsequent exercise, and students with initially poor accuracy improved in their accuracy. Students are able to evaluate themselves accurately if they are trained beforehand and given criteria and feedback to both, self-evaluation and mistakes in their solution [Tha17].

Accuracy improvement is related to experience, training, feedback and by typical standards, for example in the form of criteria [PR14, Tha17]. Appropriate feedback supports students to regulate themselves and thus improve their self-evaluation [Car20, WZH20]. In addition, criteria are required in order to acquire knowledge about the standards [Sad89, WZH20]. However, it remains open what influence usage patterns have on the accuracy of self-assessment. Usage and task design have not yet been sufficiently investigated in connection with self-assessments. Our work contributes to better understand a learner's needs and to provide better learning support through interventions.

3 The Self-Assessment question type in a nutshell

In order to support iterative improvement, the SelfAssess-plugin [Ste21], that allows students to create and assess solutions unlimited times, was implemented in Moodle. Figure 1 shows the process of working on a self-assessment question. Students can open a question to start work. The plugin shows the question and asks the student to create and upload a solution. Solutions may be uploaded as image (e.g. photo, scan) or PDF file.

Thereafter, the plugin presents a list of instructor-defined assessment criteria, offers a link to the sample solution, and asks to assess the uploaded solution. The plugin then presents feedback to the students based on their self-assessment. For this purpose, instructors defined an error tree for each question that maps typical errors to detect one or a combination of not considered criteria on one feedback text. Feedback texts help the student to improve their solution, e.g. by providing learning goal feedback (such as misconceived concepts) and learning process feedback (such as links to relevant resources, things to improve, or activities to do), until they self-assess their solution as correct or as good enough [Haa20]. After receiving the feedback, students could either improve their solution by performing a new iteration of the process (create, upload, self-assess the improved solution again, receive feedback, and take or decline another iteration) or finish the exercise.

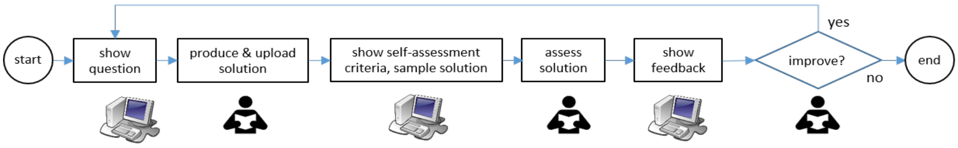
start
show question
produce & upload solution
show self-assessment criteria, sample solution
assess solution
show feedback
improve?
yes
no
end

Figure 1: Iterative problem-solving process in the SelfAssess-plugin

4 Study

In order to evaluate whether such self-assessment is practically useful, we examine in this paper three research questions. RQ1: How do students make use of voluntary self-assessment questions? Here, we examine the extent of usage over time and of questions related to each course unit, successive improvements, as well as handwritten vs. machine edited. RQ2: How well do learners self-assess their solutions? Here, we examine (1) accuracy of self-assessment by comparing student self-assessment with instructor assessment, (2) whether accuracy is related to the quality of the student answer, and (3) how accuracy of a student changes with successive trials. RQ3: What self-assessment question-related characteristics influence the self-assessment? Here, we look at the impact of item difficulty and discrimination index on self-assessment accuracy.

4.1 Method

Participants and design: The study was conducted in the compulsory course “Operating Systems and Computer Networks” of the distance learning B.Sc. Computer Science (CS course) in the winter semester 2020/2021. For the enrolled students a supplementary course was set up in a Moodle learning environment including the SelfAssess-plugin (cf. section 3). The use of the learning environment was voluntary, but conditional on a two-step consent to use the platform and to participate in the study. The second informed consent for participation in the study could be withdrawn or granted again at any time, while the first consent was required for GDPR compliance. As an incentive for students' participation, additional exercises such as self-tests, self-assessments, and assignments were offered, as well as additional tools for semester planning and for reading the digital course texts. These differences in the learning offer are comparable to different didactic offers of tutors in face-to-face teaching. Students not participating in the study had no disadvantages regarding the examination, since the course texts provided to all enrolled students form the basis of the examination.

180 of the 534 CS course participants agreed to take part in the study and to use the set up Moodle instance. By the end of the semester, the same number of active participants had been recorded. The participating students were between 19 and 65 years old (M=37.21, SD=9.03). The gender of the participants was: 128 male and 52 female.

Material: The moodle course contained four units including course texts, a usenet forum, recordings of live sessions, and questions for exam preparation. For this study, the instructor created 42 self-assessment questions (cf. dataset [HMS21]) assigned to the course units, complementing 23 MC-questions, and 30 exercises corrected by a tutor. Self-assessment questions were aimed at training constructive or analytical skills. Therefore, criteria for self-assessment were tailored to test properties of a correct solution – often implicating steps of a correct argumentation or computation. Thus, a binary scale was sufficient for letting students assess whether their solution fulfills the respective criterion.

Data collection and analysis: User interactions and user inputs within the Moodle environment have been captured in the database, especially in the standard log store. Compared to other question type plugins the SelfAssess-plugin [Ste21] enables more detailed logging capabilities. Additionally, the uploaded students' solution was stored in the Moodle data folder. These files have been used for expert rating by the course instructor. The validity of the student solution, the file type and the writing style (by hand or typewritten) was classified along with the correctness of valid uploaded solutions. Results of the manual analysis were merged afterwards with the log store data.

Procedure: The course and thus the field study began on October 1st, 2020. Students were free to choose when to start with the course and when to engage with which tasks and exercises. The course ended after 182 days, with only the first 161 days considered in this study. Results of the individual oral exams are expected within the following twelve months. Thus, the following results should be considered as preliminary.

Measures: As shown in section 3, each self-assessment question defines a number N of criteria with (1) an associated text expressing the condition that the student needs to evaluate as being fulfilled by his/her solution and (2) a boolean indicator, whether this criterion should hold for a correct solution or not (i.e. indicating a distractor). Students could open the question in the plugin, read the question text, create a solution, upload it, read the assessment criteria texts and select those that they deem fulfilled by their solution. The resulting self-assessment can be represented as a N-tuple of zeros or ones, indicating whether the i-th criterion has been selected by the student.

In order to measure the accuracy of a student's self-assessment for a given self-assessment question, we asked the instructor who defined the self-assessment question to provide a coding schema that defines when a solution fulfills a given criterion or not. Using this coding schema, the instructor rated the uploaded student solutions on the given criteria, resulting in instructor assessments represented by similar N-tuples.

The Hamming distance [Ham50] denotes the number of differences between the two N-tuples. Zero differences denote identical assessments and thus optimal accuracy. A value of N indicates completely different assessments and thus a completely inaccurate self-assessment. Student self-assessment accuracy is then defined as the Hamming similarity between the instructor assessment x and the student's self-assessment y [Ham50]:

$$Accuracy(x,y) = HammingSimilarity(x,y) = 1 - \frac{HammingDistance(x,y)}{N} \in [0..1] \subset \mathbb{R}$$

A value of zero denotes completely inaccurate self-assessment while a value of one denotes completely accurate self-assessment. This definition allows us to compare the self-assessment accuracy of questions having a different number of criteria.

While the previous definitions apply to individual students on an individual self-assessment we can extend the analysis to *question accuracy* representing the mean of all accuracies of student self-assessments of a question.

In addition, the instructor ratings enabled an evaluation of the question items using classical test theory (CTT) [DeC10]. In CTT, the item difficulty and the item discrimination index are commonly used. Item difficulty is the mean score for an item within a population of participants in a range of zero and one. For ease of interpretation, we transpose item difficulty to the difference between one and the mean score. This means that tasks with a low mean score are considered difficult, and vice versa. Since questions that are too difficult (>.7) impair motivation, the majority of questions should have a difficulty between .3 and .7. The discrimination index is the correlation of the total scores and the achieved scores for the particular task. Values above .3 are considered good, between .2 and .3 acceptable, between .1 and .2 marginal, and below .1 poor.

4.2 Results

RQ1 - Use of voluntary self-assessment questions: 131 participants performed at least on self-assessment. 97 participants uploaded only appropriate solutions regarding the question, while 7 participants consistently submitted unrelated files and 24 participants uploaded both. In 472 of the 1,472 uploaded files obviously no content-related connection with the respective question could be determined. The majority of the valid uploads was typed on a digital device, 144 files consisted of photos or screenshots of handwritings. As the semester progresses, a decreasing number of active participants and responses could be observed regarding all course activities (Fig. 2). 33 participants repeated individual self-assessment questions. Most repetitions took place within a time range of less than two days, only 15 participants repeated a question after more than two days.

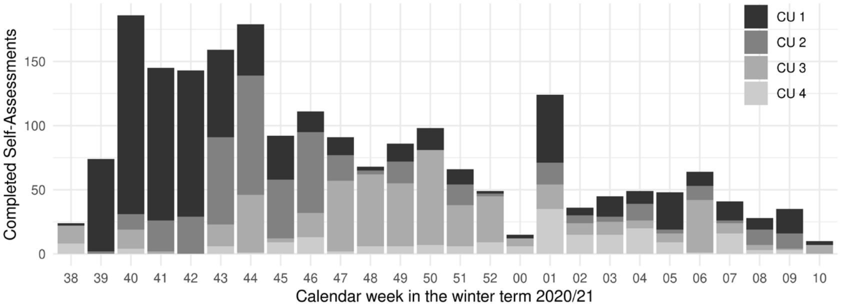

Fig. 2: Completed self-assessments for each course unit (CU) over the semester.

In order to be able to evaluate the self-assessment based on the valid uploads by an expert, it was necessary to reduce the number of observations. From the four course units we selected those questions that were completed by the most students. In course unit 4 two questions have been processed by the same number of participants, which is why we have chosen both of them. As seen in Tab. 1, five out of 43 questions [HMS21] were selected. 46 participants (35.11 % of the self-assessment users) completed at least one of the five questions. A total of 202 out of 1,000 solutions were uploaded for these five questions, but 56 of them were not related to the question and solutions of 38 more self-assessments were accidentally not recorded or canceled. Consequently, the accuracy of the self-assessment was evaluated for 108 valid uploaded solutions.

RQ2 - Accuracy of self-assessment solutions: Figure 3 shows the relationship between accuracy and the ratio of the number of fulfilled criteria, as assessed by the instructor, divided by the number of criteria N. Thus, zero achieved points denote a completely wrong answer while a one indicates a completely correct solution. Values in between indicate partially correct answers. The size of the circles indicates the number of occurrences of this combination. The color indicates the self-assessment question (task).

Obviously, students on different correctness levels showed different accuracy for a given question, whereas this relationship differs between questions. Question 262 shows that the majority of students achieve poor solutions (below .5 correctness) in combination with a relatively low accuracy (below .3). However, accuracy increases with the solution scores. Question 290 shows a diverse pattern, but the majority submitted a correct solution with good accuracy. Question 320 shows more correct solutions (above .5) with good accuracy (upper right quadrant) and a minority of weaker solutions that have been recognized as deficits by the respective participants (upper left quadrant). Questions 346 shows that all students were achieving passable solutions with higher accuracy (upper right quadrant). Question 347 shows that most students achieved few points and were also bad at assessing their work. The few solutions with higher scores showed also a higher accuracy.

In Fig. 4, the relative share for low and high accuracy and points are shown according to four quadrants for each self-assessment question. Occurrences in the top right quadrant indicate a high score and a high accuracy. Low score, but good accuracy is represented in the top left quadrant. Low values for both, score and accuracy, is mapped to the quadrant in the lower left corner, while high scores and low accuracy is shown in the quadrant at the right bottom. With respect to the population under consideration in each case, the score should span the entire range of values, so that most results fall between .4 and .8. A spread beyond this is desirable, as a differentiation of learners is necessary in the lower as well as in the upper performance spectrum. All self-assessment questions except question 346 meet this demand. Due to a comparatively high accuracy the questions 290, 320, and 346 offer the learners the chance to identify their mistakes and thus improve them based on the feedback provided by this question type.

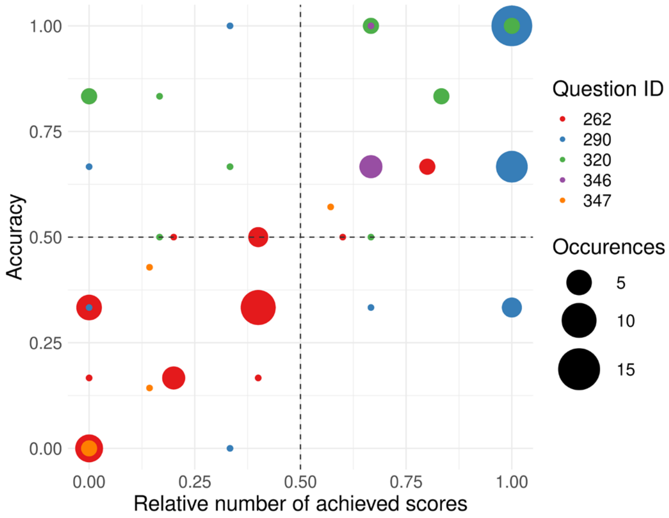

Figure 3: Self-assessment accuracy per student and question over the instructor-rated achieved points of a total of 108 edited / self-rated solutions to that question

Fig. 3 and 4 do not yet show whether questions were processed more than once and to what extent the solution was improved as a result. In 26 cases users iterated a question one more time. In 35 cases, a question was processed twice or even up to 5 times. The solution improved in 11 cases, while it worsened in two cases. In 22 cases the total score did not change, but often the self-assessment did. At the same time the self-assessment accuracy improved in 19 cases (M=.47, SD=.25), decreased in 9 cases (M=-.35, SD=.24) and remained on the same level in 7 cases.

RQ3 - Self-assessment question-related characteristics influencing the self-assessment: Following the iterative problem-solving process in Fig. 1 self-assessment question can be and characterized by three analytical dimensions: (i) question, and (ii) expected solution, (iii) self-assessment. The first dimension refers to the question itself as a starting point for the problem-solving process. The questions considered here differ qualitatively in the task design (e.g. calculation, short answers, multiple choice), proposed method complexity, scope and transfer requirements and quantitative measures like in the text length (Tab. 1) as well as indicators derived from CTT. The questions 262 and 320 have a reasonable degree of difficulty (.4-.8), while question 290 and 346 are considered easy and question 347 seems to have a very high degree of difficulty. Questions 262, 290, and 320 have a good discrimination index, thus enabling to discriminate between high-performing and low-performing students. In contrast, the questions 346 and 347 are not suitable to distinguish the performance, not least due to the low number of responses.

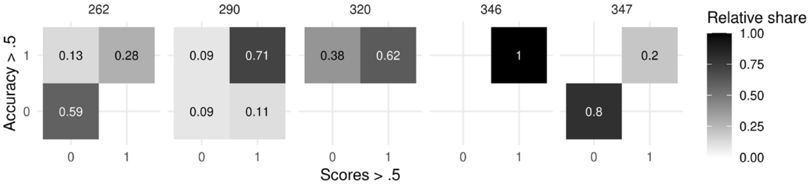

Figure 4: Relative share of occurrences for scores and accuracy below and above .5

As a second dimension the expected solution can be described by the length of the sample solution as shown in Tab. 1. The sample solutions contain between 68 and 289 words. The solution for question 347 stands out, because students are required to create 20 routing tables additionally.

ID	CU	W		SAC	N	Accuracy		Scores		τ	Idif	DI
		q	s			M	SD	M	SD			
262	1	42	97	6	46	.47	.34	.41	.38	$.70^{***}$	.59	.37
290	2	34	68	3	35	.72	.30	.82	.33	$.43^{*}$	.18	.31
320	3	77	278	6	13	.81	.18	.54	.36	.34	.46	.74
346	4	51	90	7	5	.73	.15	.67	.00	n.a.	.33	n.a.
347	4	38	289[1]	3	5	.23	.26	.17	.23	$.94^{\cdot}$	.83	.00

Tab. 1: Self-assessment accuracy per question (CU: course unit; W: text length in words; q: question; s: solution; SAC: self-assessment criteria; τ: Kendall's τ; Idiff: Item difficulty; DI: Discrimination Index; significance levels for p-values:0 (***), 0.001 (**), 0.01 (*), 0.05 (·))

As a third dimension, characteristics related to the self-assessment include the number of evaluation criteria as shown in Tab. 1 and the correspondence between students' self-evaluation and the expert rating. The evaluation of self-assessment accuracy indicated question-related differences. To better understand these differences, we computed the mean and standard deviation as well as the item difficulty and discrimination index per question (Tab. 1). While the self-assessment question in the first course unit was completed by 46 people, the two questions in course unit 4 were completed by only 5 people each (see column "N" in Tab. 1). Thus, the results become less reliable with increasing course unit number.

The mean accuracy for the questions 290, 320, and 346 show a considerable good level. Question 262 in the first course unit had most student participants, including those that did not submit solutions to the later questions. Thus, the deviation from the mean accuracy may be larger as more heterogeneous students worked on this question. The comparatively low accuracy of question 347 is related to the low number of participants and the expected extensive solution. As stated in Tab. 1 a positive correlation (Kendall's $\boldsymbol{\tau}$) between the self-assessment accuracy and the achieved points could be found.

Concerning the self-assessment accuracy, item difficulty, and discrimination index the questions 262 and 320 showed the best results. The self-assessment in 262 is a special case because the evaluation criteria included a distractor. Question 320 required a calculation with some transformations. Question 290 caused little difficulty, since the procedure required for the solution could be adopted from the course text. Question 347 stands out, considering its low mean accuracy and high item difficulty. This may be explained by the type of question.

4.3 Discussion

Our research has shown how students use self-assessment questions on a voluntary basis in a semester term (RQ1). Students engage in self-assessment questions after a short period of time, but this activity decreases as the semester proceeds and the course content progresses (see Fig. 2). This general pattern applies to all course related activities in all CS courses. Reasons may be specific to distance learning respective voluntary exercises [GGW14] or may be due to individual challenges of distance learning students who enrolled in multiple courses in parallel, underestimating the effort for learning tasks, job related demands, and seasonal effects of Christmas vacation.

Users upload both, machine-edited and handwritten solutions, indicating the need to support both in the learning process. The participants also showed ways to game the system by uploading nonsense solutions, by collecting sample solutions and self-assessment feedback. These emergent behaviors reveal usability hurdles caused by the mandatory solution uploads and the one-way test sequence imposed by Moodle. While uploads were necessary in order to conduct this research, it may have drastically reduced the total number of re-attempts and the number of participants iterating self-assessment questions immediately or with a time lag compared to past results with a plugin that did not support solution upload and display of sample solutions on demand [Haa20].

From a methodological point of view, the present study is subject to some limitations. First, only 180 out of 534 students enrolled in the course and participated in this study. The self-selection bias for the decision to learn online, to participate in a study including the provision of personal data, and then to take advantage of a specific learning offer in the form of self-assessment questions within a course could be quite large. Furthermore, we were only able to include 5 of 42 self-assessment questions in the study and had to exclude 472 unusable uploads. It is therefore possible that we were able to target particularly motivated and therefore also high-performing students. Beside that a temporary overload of the server caused the loss of a few log events that could not be stored. Because of a programming error students' self-assessment was not captured for re-attempts in 38 cases. However, limitations of this kind are not unusual for field studies. Compared to time-limited and artificially motivated lab studies, the present results show a higher practical relevance and a potential for long-term implementation.

Regarding RQ2 we investigated how well learners did self-assess their solutions. We examined accuracy of self-assessment by comparing student self-assessment with instructor assessment and investigated whether accuracy is related to the quality of the student answer. With regard to self-assessment, the values for accuracy should be as high as possible, but at least greater than .5. The consequence of inaccurate self-assessment is inappropriate feedback, which may harm iterative improvement. Our results regarding accuracy and scores (Table 1) indicate that questions 290, 320 and 346 offered students the chance to identify mistakes and improve them using the feedback.

From a methodological perspective it becomes difficult to ensure the quality of questions and self-assessment criteria that enable learners to self-assess their solutions without the collection and manual review of a student's solution. Alternative approaches to quality assurance, such as assessments by experts in the relevant field, have already been used [Haa20]. Although the comprehensibility of the evaluation criteria for these tasks was confirmed in a previous study [Haa20], the accuracy is not very informative about the completeness and individual fit of these criteria. While the instructor used a predefined binary rating schema for scoring the solutions, a second rater could be used to ensure the reliability of the scores and thus of the accuracy.

Results regarding RQ3 are distinguished between the question, the expected solution, and the self-assessment. With the help of these dimensions and the associated key figures, self-assessment questions can be described and characterized. The used measures and applied methods may be supplemented, e.g. by examining the individual chosen self-assessment criteria or by applying the Item response theory [DeC10]. By considering a greater number of different self-assessment questions it would be possible to further examine correlations between item difficulty and the accuracy of self-assessment. The indication of a positive, but not consistently highly significant, relationship between the two variables requires further investigation. However, in terms of item difficulty, a distinction must be made between the complexity and the comprehensiveness of the required solution. Our results suggest that solutions with a low or moderate extent are easier to self-assess compared to the long solutions. In addition, it can be helpful for self-assessment to have a larger number of assessment criteria including a distractor to choose from. Additionally, the process of dealing with self-assessment questions has not been examined. From the sequence of log events and the duration of individual processing steps, further insights could be gained.

5 Conclusions

Students can only make use of formative assessment, if their self-assessment is accurate enough and if they make use of feedback for improving their answer. The evaluation of self-assessment used on a voluntary basis showed that students generally try to secure feedback from exercises and feedback for themselves. The uploading of blank answers also shows that students try to circumvent the barriers of an exercise in order to still be able to retrieve the feedback information. Assessment criteria and standardized operator requirements could be better assisted by a "help" menu to prevent invalid uploads. With respect to RQ1 the results demonstrate that for quite differently designed questions, students used the degrees of freedom to upload different valid or invalid as well as handwritten or typed solutions. By comparing the self-assessment ability and the actual points scored (RQ2), we were able to identify 108 uploaded solutions that offered students the chance to identify their mistakes and improve them based on the feedback. However, the relationship between the accuracy and solution scores requires further research considering students' abilities and the difficulty of self-assessment depending on the completeness of the solution. As factors influencing the self-assessment (RQ3), we were able to identify item difficulty and, related to this, the comprehensiveness of the expected solution. Our results suggest that solutions with a low or moderate extent are easier to self-assess compared to the long solutions. In addition, fine-grained assessment criteria and distractors can aid self-assessment.

The evaluation of students' solutions has proven a suitable instrument for a practical quality assurance, which should be considered for improvement of newly introduced or random sampled self-assessment questions. Therefore, the solution input (text, formulas, images) including the upload of handwritten solutions should be streamlined for sake of usability. Overall, the self-assessment question type has shown the potential for providing scalable learning support, cross-domain applicability is an open issue. Fine-grained assessment criteria and freedom of solution input are recommended, and high usability of the plugin seems a prerequisite for increasing iterative improvement of solutions. Complexity of expected solutions is still a challenge to be explored together with earlier access to criteria in a self-assessment.

6 References

[And19] Andrade, H. L.: A critical review of research on student self-assessment. Front Educ 4:1–13, 2019.

[BLT15] Boud, D.; Lawson, R.; Thompson, D.G.: The calibration of student judgement through self-assessment: disruptive effects of assessment patterns. Higher Education Research & Development, 34 (1). S. 45-59, 2015.

[BW09] Black, P., & Wiliam, D.: Developing the theory of formative assessment. Educational Assessment, Evaluation and Accountability, 21(1), 5–13, 2009.

[Car20] Carroll, D.: Observations of student accuracy in criteria-based self-assessment. Assessment & Evaluation in Higher Education, 45:8, 1088-1105, 2020.

[DeC10] De Champlain, A.F.: A primer on classical test theory and item response theory for assessments in medical education. Medical Education, 44: 109-117, 2010.

[GGW14] Geri, N., Gafni, R., & Winer, A.: The u-curve of e-learning: course website and online video use in blended and distance learning. Interdisciplinary Journal of E-Learning and Learning Objects, 10, 1-16, 2014.

[Ham50] Hamming, R. W.: Error-detecting and error-correcting codes. In: Bell System Technical Journal, XXIX (2), S. 147–160, 1950.

[HMS21] Haake, J.M., Ma, L., Seidel, N.: Self-Assessment Questions - Operating Systems and Computer Networks, DOI: 10.5281/zenodo.5021350, 2021.

[Har17] Hartung, S.: Lernförderliches Feedback in der Online-Lehre gestalten. In H. R. Griesehop & E. Bauer (Hrsg.), Lehren und Lernen online. Springer, Wiesbaden, 199–217, 2017.

[Haa20] Haake, J. M.; Seidel, N.; Karolyi, H.; Ma, L.: Self-Assessment mit High-Information Feedback. DELFI 2020–Die 18. Fachtagung Bildungstechnologien der Gesellschaft für Informatik e.V., 2020.

[HBZ13] Hattie, J., Beywl, W., & Zierer, K.: Lernen sichtbar machen. Schneider-Verl. Hohengehren, 2013

[LEW10] Lew, M. D. N.; Alwis, W.A.M.; Schmidt, H.G.: Accuracy of students' self-assessment and their beliefs about its utility. Assessment & Evaluation in Higher Education, 35(2), 135-156, 2010.

[PR14] Panadero, E.; Romero, M.: To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. Assess. Educ. 21, 133–148, 2014.

[Sad89] Sadler, D. R.: Formative assessment and the design of instructional systems. Instructional Science, 18(2), 119–144, 1989.

[Ste21] Steinkohl, K., Burchart, M., Haake, J.M., Seidel, N.: SelfAssess Question Type Plugin for Moodle, https://github.com/D2L2/qtype_selfassess, 2021.

[TG+12] Tejeiro, R. A.; Gomez-Vallecillo, J. L.; Romero, A. F.; Pelegrina, M.; Wallace, A.; Emberley, E.: Summative self-assessment in higher education: implications of its counting towards the final mark. Electron. J. Res. Educ. Psychol. 10, 789–812, 2012.

[Tha17] Thawabieh, A. M.: A comparison between students' self-assessment and teachers' assessment. J. Curri. Teach. 6, 14–20, 2017.

[WZH20] Wisniewski, B., Zierer, K., & Hattie, J.: The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. Frontiers in Psychology, 10, 2020.