Organizing Peer Correction in Tertiary STEM Education: An Approach and its Evaluation

https://doi.org/10.3991/ijep.v9i4.10201

Niels Heller (✉), François Bry
Ludwig-Maximilian University of Munich, Munich, Germany
niels.heller@ifi.lmu.de

Abstract—This article reports on a novel higher-education course format exploiting choreographed peer reviews and self-corrections using an online learning platform. The novel course format aims to reduce teachers' workload as it was motivated by the necessity to run examinations for all courses during all terms, even though almost all courses can only be offered every second term. As a consequence, and because of a very high students-to-teacher ratio, many students have to prepare for examinations without sufficient assistance. This article describes the novel course format and reports on its evaluation in a case study. The evaluation indicates that most students benefit from the novel course format but that it is less efficient than traditional formats based on a higher teachers' involvement. The major weakness of the novel format is the insufficient participation of some students to their peer-reviewing. The article suggests and discusses possible measures to address that weakness. This article extends previous work by the same authors by providing an extended evaluation of the students' attitudes towards the course format, their behavior on the platform and their peer reviews. The course format and the learning platform are described in greater detail and the integration into related work has been expanded.

Keywords—Peer Review, Collaborative Learning, Learning Environments

1 Introduction

Today's European higher education in sciences, technology, engineering, and mathematics (STEM), is dominated by mass education, with students-to-teacher ratios of over 800 for professors, and over 70 for teaching assistants[1]. As a result, teachers do not have enough time to individually support students, be it by providing individual feedback, or in individual or classroom discussions. This poses major challenges to today's university students. Firstly, students need to have high self-regulation skills, as teachers have no time to provide regulative guidance. Secondly, STEM students often need to gather practical experience (typically by applying a learned technique to solve an exercise problem) on their own, that is, without teacher assistance. It is obvious that students struggling with a problem would largely benefit from having a

1 At the authors' faculty

teacher, or any *knowledgeable other* [39], to help them. Thirdly, students do not sufficiently benefit from classroom discussions, as they are often impeded by large class sizes [6, 28]. Open discussions are very important in STEM fields since most problems can be approached in different ways which often subtly differ. Recall that discussing alternative approaches among peers is known to be beneficial to learning [5, 21]. Finally, teachers do not have enough time to provide individual written feedback to students, in particular on written assignments. This is problematic since the quality of the feedback received greatly impacts learning. [19] Students and teachers alike have expressed their dissatisfaction with the feedback they received respectively delivered. [30]

Educational software can help to solve these problems: Self-regulation can be sustained with learning scripts (also referred to as orchestration or learning choreographies) [32], intelligent tutor and scaffolding software, and remote laboratories can provide experimental settings which allow students to gather guided hands-on experiences, social media which allow to share and discuss approaches among large numbers of students, and computer-mediated peer review which have been shown to be beneficial to learning [11,14,38] as well as they reduce the teachers' workload. [12]

Yet, learning and teaching formats that integrate these software solutions have to be designed and thoroughly evaluated. This article reports on the design and evaluation of such a learning and teaching format which has been motivated by a specific condition at the authors' university: While examinations of bachelor degree computer science courses have to be held every semester, the limited teaching staff only allow to run courses every second semester. Therefore, many students have to prepare to many examinations on their own, that is, without assistance from the teaching staff.

To cope with this situation, a new software-based learning and teaching format was devised. The format is based on self- and peer-corrections and it aims at fostering self-regulation through a well-thought choreography. The novel format is made possible by an online learning platform which supports the work-out of exercises (that are supported with direct feedback provided by the platform) and the sharing and discussing of solutions among students. The format is designed to require a minimum of teacher involvement.

Even though exploiting peer review and peer teaching in higher STEM education is promising, this has been so far rarely undertaken and therefore rarely studied. To the best of the authors' knowledge, this article is the first proposal of a course format exploiting choreographed peer reviews and self-corrections.

The format's choreography is realized as a chain of synchronized tasks such that the completion of a task by a learner (such as working out a homework assignment) usually leads to the assignment of new tasks to other learners (such as reviewing the completed assignment). An appropriate choreography is important for several reasons: It gives students precise tasks to perform, it provides common time-periods for the tasks keeping the students' learning "in phase", a pre-condition of peer teaching, and ensuring a collective experience turning a group of students into a learning community sharing common goals and therefore motivating to help each other.

At the time being, the platform is specialized on tertiary computer science and mathematics education. It provides rich interactions for these fields: Students can

write, compile, and test code in various programming languages (such as Java or Haskell), and write and verify formal mathematical proofs (such as proofs by induction). These coding and verification services are exploited in the novel course format for one student's own learning, for her review of her peers' code, and for her self-correction.

The platform can also be seen as a social medium: In addition to peer review and self-correction services, communication tools are provided which the students can use to discuss, share, and enrich the learning material provided by the teachers. A description of the platform and of its user interface is given on the project's homepage at: https://backstage2.pms.ifi.lmu.de:8080/about

The novel format has been evaluated both quantitatively and qualitatively in a bachelor's degree course, an introduction to functional programming with the programming language Haskell. In order to evaluate the learning effectiveness of the novel format, the quality of both, the homework and the peer reviews delivered by the students on the learning platform, has been assessed by human experts. Furthermore, the course attendance, that is, the participation to the assigned homework and peer review and self-correction task, as well the students' general activity on the platform have been tracked. The evaluation was guided by the following research questions:

- What is the peer reviews' quality?
- Does the quality of the reviews delivered, respectively received, by students correlate with their examination performances?
- Does solving homework assignments correlate with examination performances as it is the case with traditional course formats?
- What is the students' attitude towards the novel course format?

To answer the first research question, a simple categorizing assessment scheme of the peer reviews' quality has been worked out which assesses whether errors or the correctness in the submissions were correctly identified by the reviewers.

Using that scheme, two teaching staff members categorized independently of each other all of submitted reviews (Kohen's $\kappa = 0.85$). This evaluation revealed that 28% of the reviews were of low quality in the sense that they exhibited serious flaws. During the categorization, a number of common types of reviews emerged, which lead to a supplementary set of categories.

Surprisingly, the second question received a negative answer: Neither the quality of the reviews the students delivered, nor the quality of the reviews they received correlates with their examination performances. However, the amount of reviews students received does correlate with their examination performances what might reflect the often-observed positive correlation between doing homework and examination performances.

Investigating the third question has shown that merely submitting homework had no significant impact on examination performance. To investigate this phenomenon further, the submission quality was assessed using the platforms compilation services. This revealed that submission quality correlates positively with examination performance (Pearson's correlation coefficient, in the following Pearson's $r = 0.49$). However, this value is not statistically significant.

To answer the fourth research question, a qualitative survey has been conducted at the end of the course. The survey revealed that the general attitude of the students (completing the survey) was positive, while stressing both positive points and providing suggestions for improvements.

This article is an extended version of a conference article [20] which provides supplementary results on the system usage, student attitudes and an extended categorization of the peer reviews. The course format and the learning platform are described in greater detail in the present article than in the conference article it extends, and the section on related work has been significantly extended.

This article is structured as follows: Section 1 is this introduction, section 2 is dedicated to related work. Section 3 introduces the course format. Section 4 describes the scientific method of the case study. Section 5 reports on the results regarding the participation and attendance, the quality of peer reviews, the quality of homework submissions, and the qualitative survey. Section 6 discusses the results and makes a comparison to a previous "traditional" course. Section 7 draws a conclusion for improving course design and for further research.

2 Related Work

Peer Review: Following [31], peer review can be defined as a learning activity in which students evaluate, make judgments on, and deliver written feedback on the work of their peers.

Several meta analyses report on the learning efficiency of peer review. [11,14,38] It has to be noted that in most of the studies referenced in these meta analyses, peer review was integrated into a face-to-face teaching routine, where peer review phases were interleaved with teacher guided instruction. The teaching format presented in this article does not include such phases.

Recently, some authors have compared the positive effects of *delivering* peer reviews and *receiving* peer reviews. [10,31,27] The positive impact on learning of delivering reviews is explained by the longer time learners spend on a subject [38] and by the reflection on one's own learning triggered by reviewing the work of others. [31]

Among the benefits of peer reviewing for learners the comparison of different approaches and standard of work, a more timely feedback, and the exchange information and ideas are cited. [18,36,12,25] Among the negative impacts of peer reviews on learning, the difficulty of making accurate assessments is reported. [18] Of special interest is peer assessment, where students grade their peers' work. While peer grades are known to correlate with teacher given grades (see for instance [12,13,25]), it has been shown that students can perceive such assessments as unfair, especially if the teacher does not provide supplementary assessments. [22] For these reasons, students participating in the peer review evaluated in this course, where asked to give text-based formative reviews without assigning grades or grade-like assessments.

Peer Teaching. Peer teaching is simply defined as a form of instruction where learners teach each other. [17] Peer teaching is known to improve teamwork abilities

and social skills among learners [35] and to contribute to the learners' comprehension. [2, 35]

Like peer review, peer teaching has been shown to be beneficial both for learners acting as "teacher" and learners acting as "student" [16] which is explained by the active engagement required by both roles. [17] Peer review can be the basis for peer teaching by taking the initial review as a starting point for a dialogue between reviewer and reviewee. Indeed, it has been shown that feedback that is received in form of a dialogue is more efficient (and more satisfactory for the learners) than the "one way communication" typical to written feedback. [30]

The educational software that provided the peer review functionality for this research allows for such review dialogues, which could sustain peer teaching.

A difficulty of peer teaching is the choice of suitable peer learning groups or pairs. This difficulty can be overcome by letting instructors decide on the pairings [17] or by relying on previous achievements to form inhomogeneous groups. [9]

Skill and Models. The introduced course format required the learners to perform both learning (in the sense of skill acquisition) and peer review with minimal teacher intervention, while in the literature peer review is typically performed after a teacher guided learning phase.

The evaluation below shows that, in this study, learners' proficiency correlated positively with the quality of given reviews. This effect was *not* found in the meta analyses in. [14,35] Yet, certain theories would predict such an outcome. Fischer's Skill Theory [15] postulates that learners construct a hierarchical framework of skills where high level skills depend on lower level skills. [35] It can be argued that delivering high quality reviews on a topic is a skill that requires skills in that topic.

Several findings presented below in Section 5 indicate that learners had reached different levels in the sense of Fischer's Skill Theory.

Other theories supporting the findings are the Conscious-Competence Model of Burch that emphasizes the importance of being aware of one's own lack of competence in early phases of learning [7], Newman's Hierarchical Error Model [29] which is based on a hierarchy of steps (which may result in errors) in problem solving tasks, and the Kruger Dunning effect that is often quoted with the following phrase: "we argue that the skills that engender competence in a particular domain are often the very same skills necessary to evaluate competence in that domain." [26 p.1]

3 Learning Platform and Course Format

The course format proposed in this article is based on three types of tasks: weekly homework, review, and self-correction (or re-work). These tasks are choreographed:

- At the beginning of a topic's first week, the topic's course material and corresponding homework assignments are published on the learning platform. The students have to deliver their homework within that first week.
- At the beginning of a topic's second week, that is, after the students delivered their topic's homework, each student is tasked to review the homework of two other students. The students have to deliver their reviews within that second week.

- At the beginning of a topic's third week, that is, after the students delivered their reviews, "blue prints" or exemplary solutions for the homework assignments are published on the learning platform. The students have to deliver corrections of both their own homework and of the two peer reviews they delivered.

The third phase is introduced to exploit the beneficial effects of self-correction on learning. [33]

While a topic is learned over three weeks, every week a new topic is introduced, that is, the aforementioned three successive one-week-long phases of a topic overlap with that of other topics. In other words, with the third week of a course, a student learns a course's topic, reviews the homework of two other students referring to the previous course's topic, and performs a self-correction of her own homework and own reviews referring to course's second to last topic. This interleaved scheme has been selected so as to exploit the positive impact of timely spaced instruction [37], and shuffled instruction. [34]

Figure 1 shows the user interface of the platform used for reviewing, displaying data taken from the course: On top, the submission to review is displayed, below two reviews (in the form of comments) were attached by two peers.

The learning platform is specifically tuned to the novel course format as it supports, among others, the aforementioned multi-phased choreography requiring almost no supervision, thus freeing the teaching staff from time-consuming “administrative” or “organizational” chores. On the platform, learning activities are organized in projects. A project encompasses teachers, learners, documents, and assignments. Any member of a project can add documents to the project and comment on the documents of all users. Documents fulfill several purposes: they can contain supplementary material, general questions, exercises (in the sense of problem definitions), or student submissions. In addition, PDF and plain text Documents, a project can also encompass code documents (as seen in Figure 1) containing source code which can be run by all users. Finally, exercise assignments ask users to provide a new document (in accordance to a given exercise document) and review assignments task users to leave a review comment on a document.

4 Evaluation Method

To evaluate the course format, an introductory course on functional programming using the programming language Haskell was conducted in the winter term of 2018 at the Ludwig-Maximilian University of Munich.

Participants. 45 students enrolled in bachelor computer science programmes of whom 12 were female and 34 males attended the course. The students were studying in their second to eighth semester.

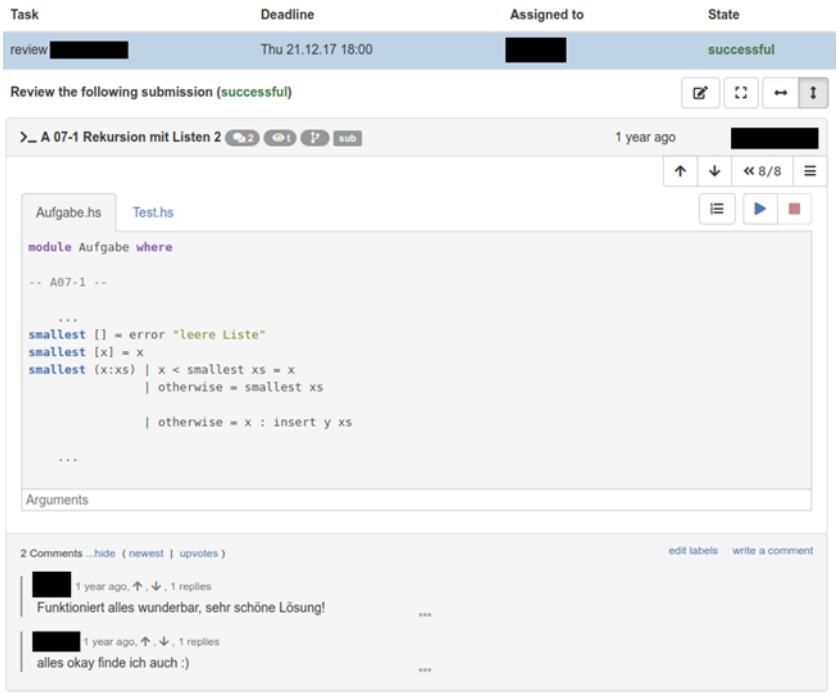

Fig. 1. Screen-shot of the reviewing interface of the platform, with redacted usernames. Top: the submitted homework, bottom: two peer reviews.

Procedure. The course lasted 13 weeks, covered 11 topics that were worked out by the students according to the scheme described in Section 3. A topic's homework encompassed two or three exercises, consisting of either a programming task or a set of questions. In total, 27 exercise solutions could be delivered by every student.

Students who missed three consecutive deadlines for a homework delivery or a peer review delivery were removed from the course on the grounds that contributing to the course, both by delivering one's own homework for other students to review and by reviewing the homework of others, is necessary for peers to learn well. This rule was made clear before registration to the course and was accepted by all students who registered for the course. After a student missed two consecutive deadlines, a warning email was automatically sent to the student by the choreography component of the learning platform.

Dataset. After the course, the quality of all student submissions, homework, and peer reviews was assessed both by members of the teaching staff and by software specifically designed for this purpose. This quality assessment was performed only for the evaluation reported in this article, its human-performed component is not part of the course format. Its automated component is part of the course format that provides immediate feedback to students.

In order to assess the peer reviews' quality, each review was categorized by members of the teaching staff after the following scheme:

+FF: "false correctly reported by the reviewer as false"
-FC: "false wrongly reported by the reviewer as correct"
+CC: "correct correctly reported by the reviewer as correct"
-CF: "correct wrongly reported by the reviewer as false"

The correctness of program submissions was assessed automatically using the standard Haskell compiler [22] and by running pre-defined unit tests, that is, tests that compare expected and computed results for a set of inputs. This way, programming submissions were categorized according to the following scheme: "wrong format" for submissions that were text of PDF files but no Haskell programs, "not compiling " for submissions the compilations of which failed (usually because of syntax errors), "compiling with failed tests" for submissions that compiled (without errors) but failed unit tests, and "tests passed" for submissions that compiled and passed the unit tests, hence that could be considered correct.[2]

These four categories can be considered as steps that have to be consecutively mastered by learners. Indeed, for beginners, the first obstacle to coding is to select the appropriate format, the second obstacle is to write code that compiles (without errors), and the third obstacle is writing code that passes the unit tests. Thus, the automatic categorization scheme reflects levels of skills as proposed by Fischer's skill theory. [15]

The students' learning behavior during the course was assessed as the number of homework and reviews they delivered and when they delivered it.

After the course, an examination referring to the course's topics took place. After that examination, a qualitative survey was conducted to assess the students' attitude towards the novel course format, the learning platform supporting it, as well as the student perception of the course format's usefulness for learning. 18 students who had attended the course and took the course's examination completed that survey.

Of the 45 students, who attended the course, 32 took the course's examination. These students' data forms the dataset of the evaluation this article reports about.

5 Evaluation Results

Participation, Drop Out and System Usage. Throughout the course, students dropped out. Most of them were removed in application of the rule mentioned at the beginning of Section 4 after they missed three consecutive deadlines. Two students freely chose to leave the course after the third week. Figure 2 illustrates the decline of the participation, notably after the third, sixth, and ninth week.

In average, students spent 2.9 hours per week on the platform, with the maximal weekly activity time of 23 hours (standard deviation 3.8 hours, median 2 hours). This confirms data gathered in the survey, where most students indicated to have spent 2-5 hours on the platform per week (choices were: "not at all", "less than 2 hours", "2-5 hours", "4-5 hours", "5 hours – 24 hours", "more than 24 hours").

2 This assumption is reasonable for the short Haskell programs beginners can write.

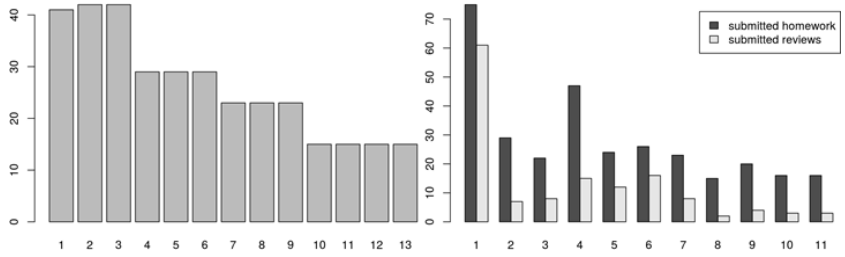

Fig. 2. Left: Numbers of students at each week, Right: Numbers of homework and peer reviews submissions at each week

All students indicated that they had spent most of the time online solving homework assignments, with reviewing course material being the second, and giving peer review being the third most frequent activity for most students.

Homework and Peer Reviews Delivered. In total 316 homework submissions and 147 reviews were delivered. Similar to course participation, the number of submitted homework and reviews declined throughout the course, as shown by Figure 2. After the first week, only a fraction of the homework submissions were peer reviewed. Peer review participation varied largely, with 18 of 45 students giving 90% of the reviews. On the platform, every comment on a document (such as a review of a homework submission) can receive multiple replies, which in turn can be replied on, possibly yielding a conversation tree. As discussed in section 2, such a functionality might enrich the peer review process, enabling a longer-lasting form of peer teaching.

Of all 147 reviews, only 15% received one or more replies by the reviewee, and only two had more replies after that.

Reviews were relatively short: half of all reviews contained 14 words or less, with outliers: 9 reviews contained 200 words or more with one containing 657 words. 69 reviews contained some kind of exemplary code.

Peer Review Quality. An evaluation of the quality assessment of the peer reviews described in Section 4 reveals that most reviews were correct in the sense that they correctly identified either errors or correctness.

The relative frequencies of labels is as follows:

+FF: 25% (“false correctly reported by the reviewer as false”)
-FC: 22% (“false wrongly reported by the reviewer as correct”)
+CC: 47% (“correct correctly reported by the reviewer as correct”)
-CF: 6% (“correct wrongly reported by the reviewer as false”)

Interestingly, only 6% of the reviews identified errors where they were none and 22% failed to indicate errors.

The correlations between the frequencies of the labels +CC and +FF where significantly positive (Pearson's r = 0.44, p=0.05) and the frequencies of the labels +FF and -FC significantly negative (Pearson's r = -0.45, p = 0.03). Other correlations between the frequencies of the labels were not significant. This suggests that students good at

spotting errors of their peers are also good at identifying correct submissions of their peers and therefore are little prone to give false feedback.

To estimate a student's average review quality, for each student a review score defined as the relative frequency of the number of correct reviews (+CC and +FF) minus the relative frequency of the number incorrect reviews (-CF and -FC) has been computed. The review scores correlate positively with the relative frequency of the number of peer reviews delivered ($r = 0.4$, $p = 0.05$), indicating that good reviewers (in the sense of delivering quality reviews) are more likely to deliver their peer reviews.

During the categorization process, common types of both flawed and accurate reviews became evident. Common types of flawed reviews were: misleading remarks, such as suggestions to solve a problem that would not solve the problem (4%), mere alternative solutions without further comments (14%), and erroneous alternative solutions (9%). On the other hand, types of accurate reviews could also be found: 2% of all reviews referred to helpful resources and 14% gave detailed explanations that were largely or completely correct.

Finally, 10% of the reviews contained admittances of the reviewers indicating that they did not fully understand the exercise and therefore could not provide a sensible review.

Although the participation in peer reviews was low, those students receiving reviews profited from them: Indeed, the relative frequency of the number of received reviews per homework submission correlates positively with the examination performance

($r = 0.44$, $p = 0.03$).

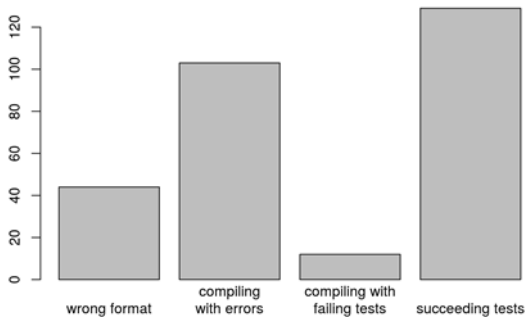

Fig. 3. Numbers of code submissions in the respective categories

Homework Quality. Of the 316 homework submissions, 232 contained executable code files. The remaining 84 homework submissions either referred to non-coding assignments (40 submissions) or were erroneously submitted in a wrong format (like Word or PDF, 44 submissions).

Of the 232 code submissions, only 129 compiled (without errors). Most of the non-compiling submissions contained syntax errors. Interestingly, of the 129 compiling submissions, only 12 failed to pass the unit tests suggesting that the automatic testing approach makes sense for such a course.

Considering the "problem solution steps" mentioned in Section 4 reveals that most students failed during the first two steps while the last step did not seem much of a hurdle for those students who mastered the previous steps. This is remarkable because it is in the last step (writing code that passes the unit test) that the actual problem is solved. The total frequencies are shown on Figure 3.

The number of submissions compiling (without errors) of a student correlates positively with the relative frequency of the number of peer reviews that student delivered (Pearson's $r = 0.35$, $p = 0.01$). This indicates that students able to solve the programming assignments are more likely to deliver peer reviews.

The number of submissions compiling (without errors) also correlates with the examination results ($r = 0.44$) but this value is not significant.

Students' Attitudes Towards Peer Review. The perceived usefulness of both delivering and receiving peer reviews was assessed. Most students (44%) indicated that delivering peer reviews was "mostly helpful" for their learning, while on the other hand, most students indicated that *receiving* peer reviews was only sometimes useful.

While the received peer reviews are rarely experienced as helpful, the students are relatively confident that their reviews were useful to others (median of 4, on a 6 point Likert scale ranging from "not useful at all" to "absolutely useful").

Students mentioned advantages of the course's peer reviews: One student stated, that topics are "learned twice", once while working on assignments, and once while peer reviewing. The opportunity to see and think about different solutions, to learn from one's peers, and to compare homework standards was also remarked. Worthwhile noting is the comment: "Peer review gave me evidence that I'm not the only one too stupid to understand the topic."

Weaknesses of the peer reviews were also mentioned: the low number of reviews received, and the low quality of some reviews.

Figure 4 illustrates the perceived usefulness of receiving and delivering peer reviews.

Students' Attitudes Towards Provided Material and Functions. The course's learning material and homework exercises were perceived as very useful for learning (median of 5 on 6-point Likert scale ranging from "not useful at all" to "completely useful"). The online compiler and the unit tests were also perceived as useful (median of 3.5 and 4.5 on the same scale).

Further, the students were asked to state general positive and negative aspects of the platform and the format. Positive aspects included notifications that indicated the current number of online learners, the simple design and structure of the platform, the possibilities to easily establish contact to tutors and peers, and the easy handling of the functionalities.

Negative aspects included the complex User Interface and the steep learning curve, and the lack of face to face meetings.

Reasons for Drop Out. Students were also asked if they dropped out of the course, and, if so, why. The reasons given were personal reasons like time constraints, and loss of motivation due to a too small number of received reviews.

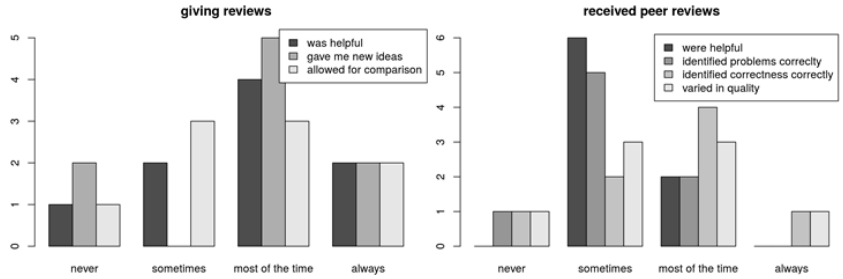

Fig. 4. Perceived properties of given and received peer reviews

6 Discussion

Peer Review Quality. The fact that the average quality of the received peer reviews did not correlate with the examination performances is surprising since the importance of the feedback quality for learning has often been stressed in the literature. [19] This surprising fact might be explained by the small number (32) of students completing the course's examination or by the emphasis of the course format on self-learning: reading low quality reviews might motivate to learn more.

The evaluation revealed that half of the reviews contained 14 words or less. To put this number into context, tutor comments of a previous course on the same topic were examined, revealing that here, half of tutor reviews were of 20 words or less, which is not much longer.

Some students indicated in their review, that they could not give a sensible review, as they did not fully understand the exercise. Interestingly, the number of these reviews was much smaller than the number of flawed reviews (10% compared to 28%), which might indicate that many of the students that did not understand the exercise, were unaware that they did not understand the exercise.

Homework and Examination Performance. The number of homework submissions did not correlate significantly with the examination performances. As a comparison, data from a previous course on the same topic was examined. That preceding course was held with a teaching staff consisting of 10 tutors who reviewed all homework submissions and a professor who hold lectures once a week. It included neither peer reviews nor self-correction. 593 students, of which 419 attended the final examination, attended the course. The lecture material and exercises were, except for minor changes, the same in both courses.

Figure 5 shows the relation between examination performance and homework submissions in the previous course. Two observations can be drawn from the figure: Firstly, students who submitted no homework do not necessarily fail in the examination. In fact, these students achieved an average mark of 64%. Secondly, submitting enough homework was a sufficient, but non-necessary, condition for examination success, as the almost empty bottom-right triangle of Figure 5 shows.

In the novel course in contrast, submitting enough homework was not a sufficient condition for examination success, indicating that the novel format helped students

struggling with the course's content less in overcoming their learning problems than the previous course did.

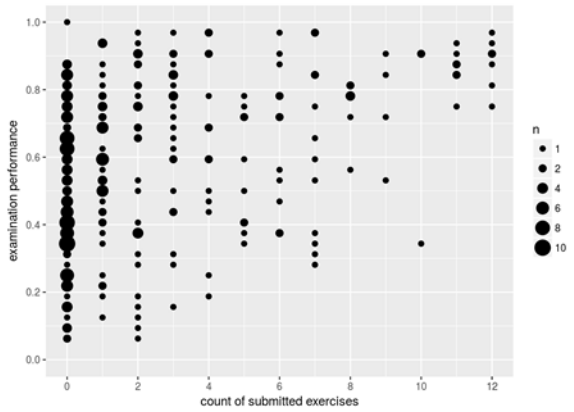

Fig. 5. Relation of examination performance to number of submitted homeworks (aggregated by week) in the preceding course

7 Conclusion

Improving the Course Format. The evaluation identified low participation in the peer review process and low review quality as a shortcoming of the course format which may be caused by the novelty of the format, requiring tasks the students are not accustomed to. While the results do not allow to conclude that the peer-review method was beneficial for the students' learning, three measures can be envisaged to improve the course format with the goal to provide all students with constant and possibly better peer reviews:

- Rather than pairing students randomly, proficient students (who are more likely to provide good reviews) could review "struggling students" (who would benefit most from having their homework reviewed) and vice versa. This would increase the reviewing efficiency without increasing the teachers' involvement. To identify proficient students, the four submission categories of Section 4 could be used. This approach would provide a very natural pairing: Those who are able to write syntactically correct code should be able to help those struggling with that task.
- Peer review quality could be improved by providing a review scheme, as sort of conceptual scaffolding. [23] Again, the submission categories of Section 3 could be used in asking questions like "Does this submission contain valid Haskell code?" or "Does it compile?"
- Finally, the social dimension of the course design could be improved: While the platform offered possibilities to reply on a received review in order to start a discussion, this functionality was only rarely used. The platform could encourage such behavior: In the case of a missing or unclear peer review, reviewees could be pro-

vided with means to contact their reviewers directly. This could naturally change the course format from a fixed three-step script (submission, review, rework) to a personalized design where the process of working out a problem, discussing solutions, and reworking solutions takes as many steps as needed.

The proposed improvements rely in part on the discovery of submission categories which in turn relied on automated compiling and testing. This seems to make the use of these techniques in other (non-programming) courses impractical. However, it can be argued that STEM subjects are often expressed in formal languages (such as algebraic expressions in mathematics or structural formulas in chemistry). The evaluated platform already supports a small set of formal languages taken from the field of discrete mathematics. Arguably, novel course formats requiring less teacher involvement could benefit from such techniques, especially in STEM education, since these techniques not only identify *proficient*, but also *motivated* students.

This article has introduced a novel course format which requires a minimal involvement of teachers. The course format has been evaluated in a case study during a university course in computer science. The proposed format relies on peer reviews and self-correction. The evaluation has shown the effectiveness of the approach and that an insufficient participation in peer reviewing, and hence a lack of reviews, was a problem. Perspectives for overcoming this problem without requiring more teacher work and for applying the format to other subjects have been discussed.

8 Acknowledgement

The authors are thankful to Elisabeth Lempa for her contribution to assessing the quality of reviews and to coding.

9 References

[1] Adams, L. (2011). Learning a new skill is easier said than done. Gordon Training International.

[2] Bathini, P. P., & Sen, S. (2017). Impact of integration through peer instructed lectures. International Journal of Basic & Clinical Pharmacology, 6(6), 1293-1296. https://doi.org/10.18203/2319-2003.ijbcp20172045

[3] Benè, K. L., & Bergus, G. (2014). When learners become teachers: a review of peer teaching in medical student education. Family medicine, 46(10), 783-787.

[4] Bester, L., Muller, G., Munge, B., Morse, M., & Meyers, N. (2017). Those who teach learn: Near-peer teaching as outdoor environmental education curriculum and pedagogy. Journal of Outdoor and Environmental Education, 20(1), 35-46. https://doi.org/10.1007/bf03401001

[5] Bonwell, C. C., & Eison, J. A. (1991). Active Learning: Creating Excitement in the Classroom. ERIC Digest.

[6] Boyle, J. T., & Nicol, D. J. (2003). Using classroom communication systems to support interaction and discussion in large class settings. ALT-J, 11(3), 43-57. https://doi.org/10.1080/0968776030110305

[7] Burch, N. (1970). The four stages for learning any new skill. Gordon Training International, CA.
[8] Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. Review of educational research, 65(3), 245-281. https://doi.org/10.3102/00346543065003245
[9] Carrell, S. E., Sacerdote, B. I., & West, J. E. (2013). From natural variation to optimal policy? The importance of endogenous peer group formation. Econometrica, 81(3), 855-882. https://doi.org/10.3982/ecta10168
[10] Cho, K., & MacArthur, C. (2011). Learning by reviewing. Journal of Educational Psychology, 103(1), 73.
[11] Dochy, F. J. R. C., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. Studies in Higher education, 24(3), 331-350. https://doi.org/10.1080/03075079912331379935
[12] Dolezal, D., Posekany, A., Roschger, C., Koppensteiner, G., Motschnig, R. and Pucher, R., 2018. Person-centered learning using peer review method–an evaluation and a concept for student-centered classrooms. International Journal of Engineering Pedagogy, 8(1), pp.1 - 147. https://doi.org/10.3991/ijep.v8i1.8099
[13] Dominguez, C., da Silva Nascimento, M., Maia, A., Pedrosa, D., & Cruz, G. (2014). Come together: peer review with energy engineering students. International Journal of Engineering Pedagogy (iJEP), 4(5), 34-41. https://doi.org/10.3991/ijep.v4i5.3537
[14] Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. Review of educational research, 70(3), 287-322. https://doi.org/10.3102/00346543070003287
[15] Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. Psychological review, 87(6), 477. https://doi.org/10.1037/0033-295x.87.6.477
[16] Gartner, A. (1971). Children teach children: Learning by teaching.
[17] Goldschmid, B., & Goldschmid, M. L. (1976). Peer teaching in higher education: A review. Higher Education, 5(1), 9-33. https://doi.org/10.1007/bf01677204
[18] Hanrahan, S. J., & Isaacs, G. (2001). Assessing self-and peer-assessment: The students' views. Higher Education Research & Development, 20(1), 53-70. https://doi.org/10.1080/07294360123776
[19] Hattie, J., & Timperley, H. (2007). The power of feedback. Review of educational research, 77(1), 81-112. https://doi.org/10.3102/003465430298487
[20] Heller, N. and Bry, F. (25-28 September 2018). Peer teaching in tertiary stem education: A case study. In The Challenges of the Digital Transformation in Education - Proceedings of the 21st International Conference on Interactive Collaborative Learning (ICL2018), volume 2. Springer. https://doi.org/10.1007/978-3-030-11932-4_9
[21] Inagaki, K., Hatano, G., & Morita, E. (1998). Construction of mathematical knowledge through whole-class discussion. Learning and Instruction, 8(6), 503-526. https://doi.org/10.1016/s0959-4752(98)00032-2
[22] Jones, S. P., Hall, C., Hammond, K., Partain, W., & Wadler, P. (1993, July). The Glasgow Haskell compiler: a technical overview. In Proc. UK Joint Framework for Information Technology (JFIT) Technical Conference (Vol. 93). https://doi.org/10.1007/978-1-4471-3215-8_6
[23] Jumaat, N. F., & Tasir, Z. (2014, April). Instructional scaffolding in online learning environment: A meta-analysis. In Teaching and Learning in Computing and Engineering (LaTiCE), 2014 International Conference on (pp. 74-77). IEEE. https://doi.org/10.1109/latice.2014.22

[24] Kaufman, J. H., & Schunn, C. D. (2011). Students' perceptions about peer assessment for writing: their origin and impact on revision work. Instructional Science, 39(3), 387-406. https://doi.org/10.1007/s11251-010-9133-6
[25] Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of personality and social psychology, 77(6), 1121. https://doi.org/10.1037//0022-3514.77.6.1121
[26] Larkin, T., 2014. The student conference: A model of authentic assessment. International Journal of Engineering Pedagogy, 4(2), pp.36-46.
[27] Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. Journal of second language writing, 18(1), 30-43. https://doi.org/10.1016/j.jslw.2008.06.002
[28] Mulryan-Kyne, C. (2010). Teaching large classes at college and university level: Challenges and opportunities. Teaching in Higher Education, 15(2), 175-185. https://doi.org/10.1080/13562511003620001
[29] Newman, M. A. (1977). An analysis of sixth-grade pupil's error on written mathematical tasks. Victorian Institute for Educational Research Bulletin, 39, 31-43.
[30] Nicol, D. (2010). From monologue to dialogue: improving written feedback processes in mass higher education. Assessment & Evaluation in Higher Education, 35(5), 501-517. https://doi.org/10.1080/02602931003786559
[31] Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: a peer review perspective. Assessment & Evaluation in Higher Education, 39(1), 102-122. https://doi.org/10.1080/02602938.2013.795518
[32] Panadero, E., Tapia, J. A., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. Learning and individual differences, 22(6), 806-813. https://doi.org/10.1016/j.lindif.2012.04.007
[33] Ramdass, D., & Zimmerman, B. J. (2008). Effects of self-correction strategy training on middle school students' self-efficacy, self-evaluation, and mathematics division learning. Journal of advanced academics, 20(1), 18-41. https://doi.org/10.4219/jaa-2008-869
[34] Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. Instructional Science, 35(6), 481-498. https://doi.org/10.1007/s11251-007-9015-8
[35] Schwartz, M. S., & Fischer, K. W. (2003). Building vs. borrowing: The challenge of actively constructing ideas in post-secondary education. Liberal Education, 89(3), 22-29.
[36] Seenan, C., Shanmugam, S., & Stewart, J. (2016). Group peer teaching: A strategy for building confidence in communication and teamwork skills in physical therapy students. Journal of Physical Therapy Education, 30(3), 40-49. https://doi.org/10.1097/00001416-201630030-00008
[37] Shaughnessy, J. J. (1977). Long-term retention and the spacing effect in free-recall and frequency judgments. The American Journal of Psychology, 587-598. https://doi.org/10.2307/1421733
[38] Topping, K. (1998). Peer assessment between students in colleges and universities. Review of educational Research, 68(3), 249-276. https://doi.org/10.3102/00346543068003249
[39] Vygotsky, L. S. (1978). Mind in society: The development of higher mental process.
[40] Williams, E. (1992). Student attitudes towards approaches to learning and assessment. Assessment and evaluation in higher education, 17(1), 45-58. https://doi.org/10.1080/0260293920170105

10 Authors

Niels Heller, born 1987, works since 2014 as a research associate at Ludwig-Maximilian University of Munich, Germany (Oettingenstraße 67, D-80538 München). He currently works in research in Technology Enhanced Learning.

Francois Bry, born 1956, works currently in research on declarative programming, human computation, and technology enhanced learning. Formerly, he worked on knowledge representation and processing, databases, automated theorem proving, and logic programming. Since 1994, he is a full professor and head of a group at the Institute for Informatics of Ludwig-Maximilian University of Munich, Germany (Oettingenstraße 67, D-80538 München).

This article is a revised version of a paper presented at the International Conference on Interactive Collaborative Learning (ICL2018), held September 2018, in Kos, Greece. Article submitted 2019-01-24. Resubmitted 2019-04-15. Final acceptance 2019-04-23. Final version published as submitted by the authors.