Peer Teaching in Tertiary STEM Education: A Case Study

[redacted for peer review]

[redacted for peer review]

Abstract. This article reports on a novel higher-education course format exploiting choreographed peer reviews and self corrections so as to reduce to a minimum the teachers' involvement. The novel course format was motivated by the necessity to run examinations for all courses during all terms, even though almost all courses are offered only 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 much higher teachers' involvement. The major weakness of the novel format is an insufficient dedication of some students to their reviewing. The article suggests and discusses possible measures to address that weakness.

1 Introduction

This article reports on a novel format for higher-education introductory courses exploiting peer reviews and self corrections so as to reduce to a minimum the teachers' involvement. The novel course format relies on a well-thought choreography of peer and self correction so as first to provide fast feedback through peer reviews to all students and to ensure a good learning through self-correction. The novel course format was motivated by the necessity to run examinations for all courses during all terms in a bachelor course of studies in computing and related fields such am bio-informatics and media informatics, even though almost all courses are offered only every second term. As a consequence and because of students to teacher ratios of over 800 for professors and over 70 for teaching assistants, many students have to prepare for examinations without sufficient assistance.

The positive impact of peer reviews of students' homework on the learning of both reviewers and reviewees has been demonstrated in former studies [25, 9, 8]. Even though exploiting peer review in higher sciences, technology, engineering, and mathematics (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. An appropriate choreography is important for several reasons: It gives students precise tasks to perform, it provides common time periods for these tasks keeping the students' learning "in phase", a pre-condition of peer review, and ensuring a collective experience turning a group of students into a learning community sharing common goals and therefore motivating to help each other.

The novel course format proposed in this article has been tested and evaluated in an introductory computer science course for Bachelor students, an introduction to functional programming with the programming language Haskell. To this aim, a specific web-based learning platform supporting the sophisticated multi-phased choreography of peer reviews and self-correction of the proposed novel course format has been conceived and implemented. The learning platform in addition to peer review and self-correction services also provides learning material and communication tools which the students can use to perform their learning assignments, discuss the learning material among themselves, and perform their homework. The learning platform also provides coding services: Students can write, compile, and test code without leaving the learning platform. The platform's coding services are used in the novel course format both for one student's own learning, for her review of her peers' code and for her self-correction.

The evaluation of the novel course format reported about in this article is both quantitative and qualitative. The evaluation's focus was the course format's learning effectiveness. To this aim, the quality of both, the homework and the peer reviews submitted 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 has been tracked. The evaluation was guided by the following research questions:

1. What is the peer reviews' quality?
2. Does the quality of the reviews delivered, respectively received, by students correlate with their examination performances?
3. Does solving homework assignments correlate with examination performances as it is the case with traditional course formats?
4. 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. 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.

Surprisingly, the second question received a negative answer: Neither the quality of the reviews 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. Indeed, only submitted homework can be reviewed.

Investigating the third question has shown that merely submitting homeworks had no significant impact on examination performance. To investigate this phenomenon further, the submission quality was assessed using an automated testing tool. 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. That survey revealed that the general attitude of those course's students who completed the survey toward the novel course format is positive stressing both positive and negative aspects.

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

The novel course format presented in this article refers to peer review, peer teaching, and skill theory.

Peer Review. Following [19], peer review can be defined as a learning activity in which students evaluate, make judgements on, and deliver written feedback on the work of their peers. Several meta analyses report on the learning efficiency of peer review [25, 9, 8]. Recently, some authors have compared the positive effects of *delivering* peer reviews and *receiving* peer reviews [18, 7, 19]. The positive impact on learning of delivering reviews is explained by the longer time learners spend on a subject [25] and by the reflection on one's owns learning triggered by reviewing the work of others [19].

Among the benefits of peer reviewing for learners, the comparison of different approaches and standard of work and the exchange information and ideas are cited [26, 13, 23]. Among the negative impact of peer reviews on learning, the difficulty of making accurate assessments is reported [13].

Peer Teaching. Peer teaching is simply defined as a form of instruction where learners teach each other [12]. Peer teaching is known to improve teamwork abilities and social skills among learners [23] and to contribute to the learners' comprehension [1, 3, 2]. Like peer review, peer teaching is has been shown to be beneficial both for learners acting as "teacher" and learners acting as "student" [11] what is explained by the active engagement required by both roles [12]. 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 [12] or by relying on measures of previous achievements to form inhomogeneous groups [6].

Skill Theory and Related Models. In the evaluation below, learners' proficiency correlates positively with review quality, an effect *not* found in the meta analyses of [25, 9]. The Skill Theory of Fischer [10] would predict the correlation reported about below. This theory postulates that learners construct a hierarchical framework of skills where high level skills depend on lower level skills [22]. Other theories supporting the findings reported about below are the Conscious-Competence Model of Burch that emphasises the importance of being aware of one's own lack of competence in early phases of learning [4] 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" [17, p.1].

3 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 as follows for each course "topic", or chapter:

A topic is learned in three successive one week long phases:

1. 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.
2. 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.
3. 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 a self-correction phase is introduced to exploit the beneficial effects of self-correction on learning [20].

While a topic is learned over three weeks, every week a new topic is introduced, that is, the afore-mentioned three successive one week long phases of a topic overlap with that of other topics. In other words, with the thirds week of a course, a student learns the 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 topic before last. This interleaved scheme has been selected so as to exploit the positive impact of timely spaced instruction [24], and shuffled instruction [21].

The web-based learning platform specifically tuned to the novel course format supports among others the afore-mentioned multi-phased choreography requiring almost no supervision thus freeing the teaching staff from time-consuming "administrative" or "organizational" chores.

4 Evaluation Method

Participants. The course run for evaluation purposes with the novel format and was attended by 45 students enrolled in Bachelor computer science programme of whom 12 were female and 33 male. The students were studying in their second to eighth semester.

Procedure. The course lasted 13 weeks, covered 11 topics that were worked out by the students according to the scheme described in the previous Section 3. A topic's homework consisted in two or three exercises. An exercise was 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, has been assessed both by members of the teaching staff and by a software specifically designed designed for the purpose and described below. 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 and most likely positively impact on their self-regulation [5], an impact that deserves to be evaluated in future studies.

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 [15] 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.[1]

[1] This assumption is reasonable for short Haskell programs beginners can write.

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 [10].

The students' learning behaviour during the course was assessed after 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. This section reports on the participation to and the dropout from the course and on the delivery of homework and peer reviews.

Drop Out. Throughout the course, students dropped out. Most 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 1 illustrates the decline of the participation, notably after the third, sixth, and ninth weeks.

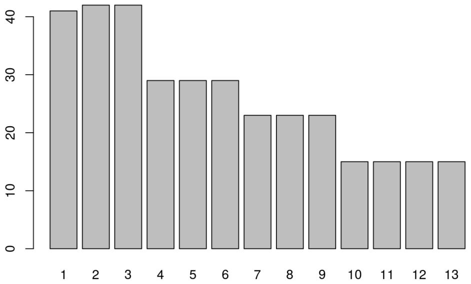

Fig. 1. Numbers of students at each week

Homework and Peer Reviews Delivered. In total 316 homeworks and 147 reviews were delivered. As with course participation, the number of submitted homework

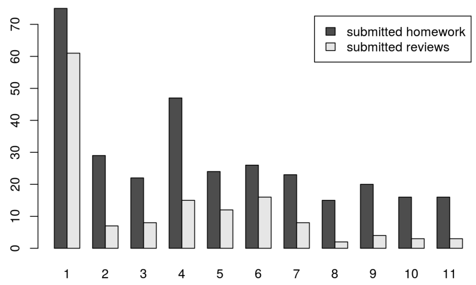

Fig. 2. Numbers of homework and peer reviews submissions at each week

and reviews declined throughout the course. Figure 2 gives the figures After the first week, only a fraction of the homeworks were peer reviewed. Peer review participation varied much, 18 of 45 students gave 90% of the peer reviews.

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 the errors of their peers are also good at identifying the correct submissions of their peers and therefore 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.

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

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.

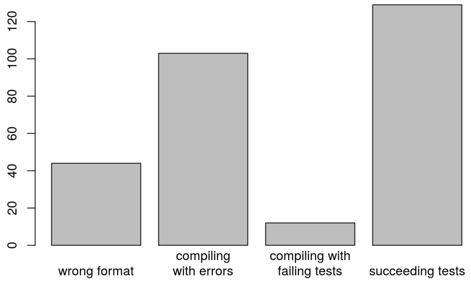

Fig. 3. Numbers of code submissions in the respective categories

Considering the "problem solution steps" mentioned in Section 4 shows 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' Attitude Towards the Novel Course Format. This section reports the results of the qualitative survey tun after the course's examination.

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

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

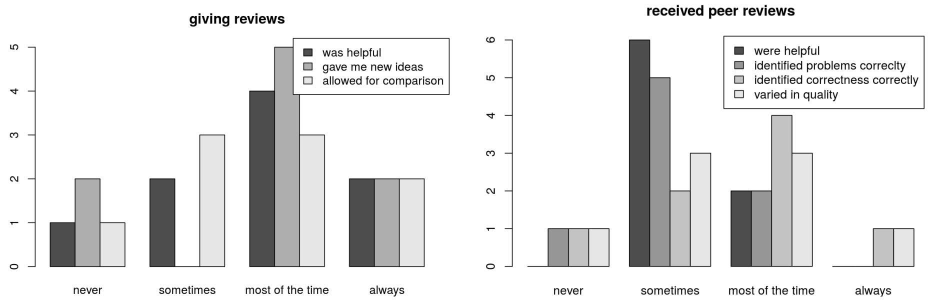

Fig. 4. Perceived properties of given and received peer reviews

Students mentioned advantages of the course's peer reviews: The opportunity to see different solutions and of learning from one's peers, and comparing homework standards. 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: low number of reviews received, and low quality of some reviews.

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).

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, loss of motivation due to a too small number of received 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 been often stressed in the literature [14]. This surprising fact can be explained as follows. Firstly, this could be due to the small number (32) of students completing the course's examination. Secondly, with the novel course format based on self-learning, reading low quality reviews might motivate to learn more. Furthermore, the students were tasked to self-correct their homework, that is, to re-work.

Homework and Examination Performance. The number of submitted homeworks does not correlate significantly with the examination performances. As a comparison, data from a previous course was examined. That preceding course was held with a teaching staff consisting of 10 tutors who reviewed all homeworks and a professor who hold lectures once a week. That preceding course format had 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.

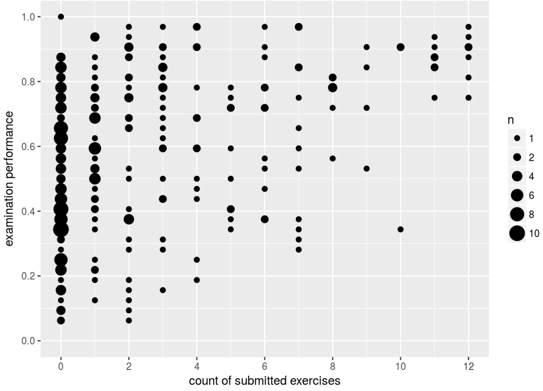

Fig. 5. Relation of examination performance to number of submitted homeworks (aggregated by week) in the preceding course

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

7 Conclusion

Improving The Course Format. As elaborated above, the first point for improvement is the provision of all students with constant and possibly better peer reviews. To this end three measures can be envisaged:

1. 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 pair-

ing: Those who are able to write syntactically correct code should be able to help those struggling with that task.

2. Peer review quality could be improved by providing a review scheme, as sort of conceptual scaffolding [16]. Again, the submission categories of Section 4 could be used in asking questions like "Does this submission contain valid Haskell code?" or "Does it compile?"
3. Finally, the social dimension of the course design could be improved. In the case of a missing or unclear peer review, the platform could provide reviewees 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 based of formal languages (such as algebraic expressions in mathematics or structural formulas in chemistry) that could be interpreted and tested automatically. It is a conviction of the authors that 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 learning 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.

Acknowledgements. The authors are thankful to [redacted for peer review] for her contribution to assessing the quality of reviews and to coding.

References

1. Bathini, P.P., Sen, S.: Impact of integration through peer instructed lectures. International Journal of Basic & Clinical Pharmacology 6(6), 1293–1296 (2017)
2. Benè, K.L., Bergus, G.: When learners become teachers: a review of peer teaching in medical student education. Family Medicine 46(10), 783–787 (2014)
3. Bester, L., Muller, G., Munge, B., Morse, M., Meyers, N.: Those who teach learn: Near-peer teaching as outdoor environmental education curriculum and pedagogy. Journal of Outdoor and Environmental Education 20(1), 35 (2017)
4. Burch, N.: The four stages for learning any new skill. Gordon Training International, CA (1970)
5. Butler, D.L., Winne, P.H.: Feedback and self-regulated learning: A theoretical synthesis. Review of Educational Research 65(3), 245–281 (1995)

6. Carrell, S.E., Sacerdote, B.I., West, J.E.: From natural variation to optimal policy? the importance of endogenous peer group formation. Econometrica 81(3), 855–882 (2013)
7. Cho, K., MacArthur, C.: Learning by reviewing. Journal of Educational Psychology 103(1), 73 (2011)
8. Dochy, F., Segers, M., Sluijsmans, D.: The use of self-, peer and co-assessment in higher education: A review. Studies in Higher Education 24(3), 331–350 (1999)
9. Falchikov, N., Goldfinch, J.: Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. Review of Educational Research 70(3), 287–322 (2000)
10. Fischer, K.W.: A theory of cognitive development: The control and construction of hierarchies of skills. Psychological Review 87(6), 477 (1980)
11. Gartner, A., et al.: Children Teach Children: Learning by Teaching. ERIC (1971)
12. Goldschmid, B., Goldschmid, M.L.: Peer teaching in higher education: a review. Higher Education 5(1), 9–33 (1976)
13. Hanrahan, S.J., Isaacs, G.: Assessing self-and peer-assessment: The students' views. Higher Education Research & Development 20(1), 53–70 (2001)
14. Hattie, J., Timperley, H.: The power of feedback. Review of Educational Research 77(1), 81–112 (2007)
15. Jones, S.P., Hall, C., Hammond, K., Partain, W., Wadler, P.: The glasgow haskell compiler: a technical overview. In: Proc. UK Joint Framework for Information Technology (JFIT) Technical Conference. vol. 93 (1993)
16. Jumaat, N.F., Tasir, Z.: 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 (2014)
17. Kruger, J., Dunning, D.: 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 (1999)
18. Lundstrom, K., Baker, W.: 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 (2009)
19. Nicol, D., Thomson, A., Breslin, C.: Rethinking feedback practices in higher education: a peer review perspective. Assessment & Evaluation in Higher Education 39(1), 102–122 (2014)
20. Ramdass, D., Zimmerman, B.J.: 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 (2008)
21. Rohrer, D., Taylor, K.: The shuffling of mathematics problems improves learning. Instructional Science 35(6), 481–498 (2007)
22. Schwartz, M.S., Fischer, K.W.: Building vs. borrowing: The challenge of actively constructing ideas in post-secondary education. Liberal Education 89(3), 22–29 (2003)
23. Seenan, C., Shanmugam, S., Stewart, J.: 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 (2016)
24. Shaughnessy, J.J.: Long-term retention and the spacing effect in free-recall and frequency judgments. The American Journal of Psychology pp. 587–598 (1977)
25. Topping, K.: Peer assessment between students in colleges and universities. Review of Educational Research 68(3), 249–276 (1998)
26. Williams, E.: Student attitudes towards approaches to learning and assessment. Assessment and Evaluation in Higher Education 17(1), 45–58 (1992)