Studies in Educational Evaluation 43 (2014) 197–205

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

The validity and students' experiences of peer assessment in a large introductory class of gene technology

Henna Asikainen [a,*], Viivi Virtanen [a], Liisa Postareff [b], Pekka Heino [a]

[a] University of Helsinki, Faculty of Biological and Environmental Sciences, Viikinkaari 9, PL 56, 00014 Helsinki, Finland
[b] University of Helsinki, Faculty of Behavioural Sciences, Siltavuorenpenger 5A, PL 9, 00014 Helsinki, Finland

ARTICLE INFO

Article history:
Received 4 April 2014
Received in revised form 25 June 2014
Accepted 7 July 2014
Available online 5 August 2014

Keywords:
Assessment practices
Peer assessment
Student evaluation
Teaching and learning
Higher education

ABSTRACT

Although studies have examined the validity of peer assessment, research including students' own experiences of peer assessment is scarce. The present study aims to improve assessment practices in a context with a highly traditional assessment culture. The aim is first to examine the validity of peer assessment by analysing the compatibility of student and teacher evaluation and explore the differences between minor and major students' evaluations. Second, the study examines students' experiences of peer assessment. Peer assessment was implemented in a large bioscience course with 79 student participants. After the peer assessment, the students provided feedback. The results indicate that student subject understanding can be supported through a proper assessment practice. Peer assessment was successful in an introductory class with minor and major students, and most students experienced it as supportive of their learning.

Introduction

In recent decades, globalization, the rapid development of communications technology and growth in knowledge has challenged higher education. Nowadays, higher education is expected to produce generic skills, such as interaction skills, information literacy reading skills, and problem-solving skills (Tynjälä, Slotte, Nieminen, Lonka, & Olkinuora, 2006). The ability to evaluate one's own skills and knowledge has also become increasingly important as requirements in working life are constantly changing. Therefore, recent research has emphasised the role of assessment in serving purposes of lifelong development (Sluijsman & Prins, 2006; Davey & Palmer, 2012). Assessing student learning involves practices that mainly serve the purpose of ranking, as well as those that in essence serve to support student learning: namely, summative and formative assessment (Brown, Bull, & Pendlebury, 1997; Bryan & Clegg, 2006; Yorke, 2003; Nicol & Macfarlane-Dick, 2006). Research has long suggested that assessment culture must change from assessment *of* learning towards assessment *for* learning, in other words, assessment should serve as a tool to monitor learning and to provide feedback for modifying learning as well as teaching (Black, Harrison, Lee, Marshall, & William, 2004; Boud & Falchikov, 2006; Stiggins, 2002).

In higher education, some have identified peer assessment as one optional assessment strategy for developing desired skills and capabilities by encouraging students to focus on constructing of knowledge and deep understanding (Somervell, 1993; Struyven, Dochy & Janssen, 2005; Lindblom-Ylänne, Pihlajamäki, & Kotkas, 2006; Davey, 2011; Davey & Palmer, 2012). Studies have shown that understanding the assessment process and criteria also helps students to evaluate their own learning (Nicol & Macfarlane-Dick, 2006). Thus, peer assessment aims to integrate learning and assessment by promoting the active engagement of learners in the assessment process, yielding better learning outcomes. In recent decades, peer assessment has been carried out in various contexts in order to promote student involvement in assessment with fruitful results (Smyth, 2004; Gouli, 2008; Welsh, 2007; Davey, 2011; McGarr & Clifford, 2013).

Despite the extensive literature on peer assessment, it is by no means self-evident that peer assessment is widely adopted into higher education teaching practices (Postareff, Virtanen, Katajavuori, & Lindblom-Ylänne, 2012; Halinen, Ruohoniemi, Katajavuori, & Virtanen, 2013). Studies by Postareff et al. (2012) and Halinen et al. (2013) suggest that assessment practices at the university level are

* Corresponding author. Tel.: +358 50 4156 715.
E-mail addresses: henna.r.asikainen@helsinki.fi (H. Asikainen), viivi.virtanen@helsinki.fi (V. Virtanen), liisa.postareff@helsinki.fi (L. Postareff), pekka.heino@helsinki.fi (P. Heino).

http://dx.doi.org/10.1016/j.stueduc.2014.07.002
0191-491X/

still quite teacher-led: the teaching and learning culture in the academic environment as well as the lack of pedagogical training create barriers to changing the tradition of teacher-led assessment. Some have stated that the culture of teacher-lead assessment influences students' engagement in peer learning: students with no experiences of peer assessment find it more difficult, and have more negative experiences of it, and require more support to adopt it (McGarr & Clifford, 2013).

One concern about the teacher-led practices also arises from recent research indicating that inappropriate assessment practices can have unwanted effects on students' study processes and achievement (Asikainen, Parpala, Virtanen, & Lindblom-Ylänne, 2013). The results implied that due to the inappropriate nature of the assessment, students applying a surface approach succeeded very well in the exam even though their qualitative self-evaluations indicated poor learning outcomes. The study presented evidence of the backwash effect of assessment, revealing how assessment practices guide student learning by having an impact on student study strategies. Teachers in the field of science are often experts in their own field of study but novices concerning pedagogical thinking (Lueddeke, 2003). The need for pedagogical training in the field of biosciences has been acknowledged (e.g., Asikainen et al., 2013; Halinen et al., 2013). However, for example in Finnish universities, pedagogical training is not a requirement for university lecturers. Thus, a lot of teachers do not participate in those courses. Alternative ways to support students' learning and change the teacher-led practices have to be developed, especially in contexts with a highly traditional assessment culture. In our study, peer assessment was implemented into a traditional lecture course with final exam by giving a minor pedagogical support for the teacher.

One reason for the lack of peer assessment practices at universities is also that peer assessment raises doubts about reliability, standards, and equity (Hinett, 1999). Reliability refers to how consistently a measurement yields similar results under varying conditions (Brown et al., 1997). Researchers have noted that peer assessment involves students in the identification of criteria and using these criteria to make judgements (Nulty, 2011). However, even criteria-based assessment has its challenges in the full complexity of multi-criterion and qualitative judgments that challenge reliability (Sadler, 2009). As Lindblom-Ylänne et al. (2006) noted, the technical aspects were graded more reliably than the content in the peer assessment of essays. A study exploring the quality criteria in peer assessment practices suggests that many of the generic quality criteria serve in peer assessment, but in an embedded way in the assessment settings (Ploegh, Tillema, & Segers, 2009).

The *validity* of assessment refers to whether an assessment meets its own intended objectives and whether the grades correspond to the quality, breadth and depth of students' academic achievement (Sadler, 2009). A meta-analysis conducted by Falchikov and Goldfinch (2000) comparing peer and teacher marks in assessment practices showed high validity in well-designed experiments with well-understood criteria. However, some studies have shown that students assign lower marks to their peers than the teacher does (Elliot & Higgins, 2005); in contrast, other results have revealed that students assign higher marks than the teacher does for descriptive questions, but mark numeric questions equally (Davey, 2011). One reason for the difference in teachers and students marks could be that students participating in lecture courses are not necessarily a homogenous group concerning their subject knowledge and competence. For example, many minor students participate in courses. A previous study has shown that the performance of minor students can be weaker than major students (Hailikari & Nevgi, 2010). Studies concerning the difference in minor and major students' competence in peer assessment have been scarce.

Previous research has shown that students have positive views of peer assessment and generally value the experience of understanding the assessment process (Ballantyne, Hughes & Mylonas 2002; Prins, Slujsmans, Kirchner, & Strijbos, 2005; Davey & Palmer, 2012; McGarr & Clifford, 2013). However, studies have shown that the validity of peer assessment does not necessarily reveal anything about the quality of students' learning outcomes (Segers & Dochy, 2001). Thus, when studying how well peer assessment works, one should also take also into account students' own experiences of it. Nevertheless, qualitative research on students' evaluations of peer assessment is scarce (McGarr & Clifford, 2013).

This study originated from the need of academics to support deep-level learning in the study of genetics. It has been shown that in the context of biosciences, the assessment culture is still highly traditional and concentrates on measuring factual knowledge (e.g, Halinen et al., 2013). A previous study has also suggested that neither the teachers nor the students always considered assessment in biosciences to be valid or reliable (Räisänen et al., 2012) and that grades do not always reflect students' learning outcomes (Räisänen et al., 2012; Asikainen et al., 2013). Furthermore, the academic staff recognised two main challenges in the teaching-learning environment: the students do not receive enough feedback on their learning, especially during their bachelor studies, and many teachers of bachelor studies are over loaded with large classes and a heavy workload. Research has shown that peer assessment is an important component of a lecturer's limited resources for providing feedback in a large class (O'Moore & Baldock, 2007; Ballantyne et al., 2002), and student involvement in assessment should also lead to better learning outcomes and a deeper understanding of the subject (Bryan & Clegg, 2006). Thus, with peer assessment, we wanted to support student learning by involving students as active participants in the assessment process. Peer assessment would offer students a learning situation closely tied to an examination, and that would serve as a tool for providing and receiving feedback.

The present study aims first to examine the validity of peer assessment by analysing the compatibility of students' and teacher's evaluations. Second, the study also aims to explore the differences in minor and major students' evaluations. Third, the study examines students' own experiences of peer assessment. Accordingly, in this study we tested our hypothesis that the students' grading would in most cases resemble the grading done by the teacher resulting in a high level of reliability between peer and teacher marking despite the students' heterogeneity in relation to their prior knowledge or motivation.Research questions are as follows:

(1) How similar are the students' and the teacher's evaluations?
(2) How similar are minor and major students' peer evaluations?
(3) How do the students experience the peer assessment?

Methodology

The study context

The gene technology course is a three credit lecture course (1 ECTS equals 27 h of work) which is offered by the genetics major at the Faculty of Biological and Environmental Sciences at the University of Helsinki. The course is obligatory for students majoring in genetics or for any university student studying a minor degree in genetics. During the seven-week course, the lecturer in genetics gives a two-hour lecture twice a week. The lectures took place in a large lecture hall where the teacher presented the subject matter to the students. About 80 students participate in the course every year. The teacher in the present course was experienced and had taught the course for several years; the students had just recently nominated the teacher as the best lecturer of the year.

Participants

A total of 79 students participated in this study. Of the 79 students, 15 were male and 64, female. The students represented diverse majors, which were grouped as follows: 1 = molecular bioscience ($N = 15$), 2 = biology ($N = 36$), 3 = geography ($N = 6$), 4 = forestry and agriculture ($N = 8$), 5 = science ($N = 8$), and, 6 = other major ($N = 3$). Three of the students did not report their major. The participants of this study represent two main groups of students: those who study genetics as their major (groups 1–2), and those who study genetics as their minor (groups 3–6).

Assessment tool

To pass the course, the students were assessed with a final exam that included three essay questions designed to focus on the core content of the course (Appendix A). We wanted to design the exam so that it would measure students' different levels of knowledge and understanding of the course content. In bioscience studies, the exams often measure how well the student can memorise or repeat factual knowledge (Halinen et al., 2013). Thus, the exam questions were formed using as a basis the model of Hailikari, Nevgi, and Lindblom-Ylänne (2007). This model is based on Bloom's revised taxonomy (Anderson, 2001) and was originally developed to measure students' prior knowledge. The model was used in formulating the exam questions and the assessment criteria to cover different levels of understanding. In addition, academic writing was also graded by assigning points for academic writing style (Table 1).

The first essay question consisted of two short questions about vector molecules and probes, asking students to describe the concepts and to provide examples of applications. This question required students of knowledge of facts and meaning, because they had to identify and understand some concepts and how they are used. The second task concerned the use of genomic and cDNA libraries for the isolation of individual genes, asking students to compare these techniques and reason. This task required understanding of these techniques and integration of knowledge by comparing and understanding the interrelations of these techniques. The third task concerned the PCR technique, asking students to describe the principle, to motivate the process, and to design the primers for a specific purpose. This question was designed to require understanding of the concepts as well as application of knowledge. The exam itself included two parts: on day one, the students wrote the exam; they had 2 h to complete the tasks. On day two, each student used the criteria offered by the teacher to peer assess two anonymous exams. If the student failed the exam or could not participate in it, he or she had other opportunities to attempt the exam later in the spring term, but without peer assessment. Student registration showed that 96 students passed the course during the spring term.

Table 1
Comparison of the average grades given by all three assessors and correlations between them..

Assessor	Teacher	Peer assessor group A	Peer assessor group B
Mean grade	2.87	2.80	2.71
Std. Deviation	1.52	1.48	1.46
Pearson's Correlations			
Teacher	1		
Peer assessor group A	0.85**	1	
Peer assessor group B	0.85**	0.73**	1

** $p > 0.001$

Three assessors (two anonymous peers and one teacher) assessed each student's final exam. Thus, our data consisted of the three separate and independent evaluations for each anonymous examination paper ($N = 79$). All the students were given an identification number which was shown on the exam paper. No other identification information, like their name, was in view. Thus, the students did not know whose exam they were assessing. The grading scale followed the universal six-level grading scale from 0 to 5 adopted by the University of Helsinki. The grading scale is directly comparable to the ECTS grading scale, and the teacher-designed maximum score for the exam was 30, which corresponds to 0 = fail or F or FX (under 13), 1 = passable or E (14–16), 2 = satisfactory or D (17–19), 3 = good or C (20–22), 4 = very good or B (23–26), and 5 = excellent or A (27–30).

Data collection

The students were given careful instructions about peer assessment. The teacher, in co-operation with the authors, designed exact criteria for the assessment (Appendix B). Thus, the students were given a criteria matrix which included descriptions of different levels of knowledge and of the point distribution. The criteria matrix also included a description of the subject content, examples of what the exam responses should include, and how to grade them, as well as the grading scale showing the relationship between the points and grades. Before starting to grade, each student read the first paper for assessment at a glance. Then, the lecturer taught the content for the first question, after which each student reread the response carefully using the criteria matrix to assess the response. The same process concerned all the questions of the first paper the students were to assess. Meanwhile, the students were able to ask clarifying questions of the lecturer. In this situation, the students actively asked questions about the core content of the course and about grading. For example, some students were unclear, what are the main principles for designing primers for PCR, core knowledge required to successfully complete task 3. The students graded each task based on the assessment matrix and gave also comments for giving those points. The comments included, for example, comment like: "the answer included some of the important points but was quite disordered and not properly outlined" or "the comprehension was clearly present and the answer was very well written". The students wrote their comments and grades on a separate paper in which they included the identification number given to the exam and their own student number while making no marks on the original response. This way the peer assessors were not able to identify whose paper they were assessing. In addition, due to the students' number, we were able to identify who had graded the exams. The students graded the second evaluation papers independently. After the peer assessment, the teacher assessed all the examinations with the same criteria matrix.

An anonymous questionnaire collected students' experiences of peer assessment right after its completion. The questionnaire requested no background information. The peer assessment was first explained as follows: 'Peer assessment offers a learning situation for both subject knowledge and academic writing, and aims to serve as a tool for receiving feedback in order to understand the criteria for one's own grade.' Next, one item measured students' experiences of peer assessment on the following scale: 0 = I cannot say, 1 = very bad, 2 = quite bad, 3 = quite good, 4 = good, and 5 = excellent. Finally, students could explain the scoring of the last item in an open-ended question.

Analyses

The correspondence between teacher's grades and those given by the students was first explored with descriptive values such as percentage and averages, and second with Pearson's correlation analyses. The differences between different assessor groups' grades were analysed with one-way ANOVA. We conducted change variables by subtracting grade B from grade A (Grade A–B) to further analyse the similarities between the assessment of the same paper by the two peer assessors A and B; this showed whether the grades were equal (0) or whether they differed from each other (±1–3). We explored the differences between the students' assessment scores according to their major were explored by examining the consistency of the grades in cases where a minor and a major student both graded the same exam.

Students' experiences of peer assessment were first examined with descriptive statistics of their responses to questions about their experiences. In addition, the students' responses to the open-ended question about their experience of peer assessment were analysed with inductive content analysis. First, the students' answers were read through several times. Second, the fragments where students described their experiences of peer assessment were separated and categorised. Students' comments were usually short and therefore easy to categorise. Students' whole responses were often categorised in a single category.

Results

Comparing the teacher's and student assessors' grading

First, we examined the descriptive values of the grades given independently by the three assessors for each exam paper. Of the 79 participants, 70 students passed the examination and 9 students failed. The average grade given by the teacher was 2.87. The average grades given by both peer assessors were quite similar (peer assessor $A = 2.80$; peer assessor $B = 2.71$). One-way ANOVA showed no significant differences between the means of the grades given by the teacher and the peer assessors ($F = 0.16$, $p = 0.85$). Correlation analyses compared the teacher's grade and the grades scored by peer assessor groups A and B. The Spearmann's rank correlation analyses revealed high correlations between the grades given by the lecturer and the peer assessors, namely, those in peer assessor groups A ($r = 0.85$, $p < 0.001$), and B ($r = 0.85$, $p < 0.001$). Accordingly, these high correlations provide evidence of agreement between peer and teacher marks. The correlation between student groups A and B was also high ($r = 0.73$, $p < 0.001$), but lower than the correlations between the students' and teacher's marks. Table 1 shows a comparison of all three assessors' grades and their correlations.

We tested the similarity of the two peer assessors' grading by computing a new variable by subtracting grade B from grade A (Grade A–B), which showed whether the grades were equal (0) or whether they differed from each other (±1–3). In most cases (89.7%), the peer marks were either the same or differed by one unit on a scale from one to five. A total of 46% of the peer marks were exactly the same. In seven cases, the marks differed from each other by two units, and in two cases, the marks differed by three units. Furthermore, we compared the teacher's grading to the grades given by the peer assessors, the similarity was even higher: the teacher's grade was mostly either the same or differed by only one unit from the grades given by assessors A or B (94.9% and 91.0%, respectively). The frequencies and percentages of the change variables appear in Table 2.

Table 2
Frequencies describing the change variables that measure differences between the grades given by the teacher and the peer assessor groups.

	Teacher—peer assessor group A		Teacher—peer assessor group B		Peer assessor Group A–Group B	
	F	%	*F*	%	*F*	%
−3					1	1.3
−2	3	3.8	2	2.6	2	2.6
−1	7	9.0	8	10.3	16	20.5
0	46	59.2	46	59.0	36	46.2
+1	21	26.9	17	21.8	17	21.8
+2	0	0	5	6.4	5	6.4
+3	1	1.3			1	1.3

Comparison of different major groups

We also examined the difference in grades between different major groups by exploring the distribution of grades in each of the major groups (1 = molecular bioscience, 2 = biology, 3 = geography, 4 = forest- and agriculture, 5 = science and 6 = other major) and also by comparing bioscience majors (molecular bioscience and biology) to bioscience minors (all the other majors). One-way ANOVA test showed no statistically significant difference ($p > 0.12$) between the major groups or major and minor groups in terms of the distribution of grades. In all the major groups, grades 3 and 4 were the most common, followed by grades 2, 5, and 1, in order of frequency.

We explored whether grades given by the minor students differed from those given by the major students and the teacher. To do this, we compared the grades given by the students and the teacher in those cases where a major and a minor student assessed the exam ($N = 25$). In 10 of those 25 cases, the grade given by the minor student was equal to the grade given by the major student and the teacher. In 11 cases, the grade given by the minor student was equal to the teacher's grade, but differed from the major student's grade. In one case, the minor student's grade was higher, and in three of the cases, the grade was lower than grades given by the teacher and the major student. These results indicate that minor students were as capable of assessing the exams as the major students were.

The marks with significant dissimilarity between the assessors

Of the 78 grades, 9 were differed between the assessors by more than one unit. The difference was similar in both lower (1–2) and upper grades (4–5). Four different distribution groups could be identified: (1) the teacher's mark is equal to the lower mark given by either of the peer assessors ($N = 2$), (2) the teacher's mark falls between the peer assessors' marks ($N = 3$), (3) teacher mark equals higher mark given by either of peer assessors ($N = 3$), and (4) the teacher's mark is higher than the marks given by the peer assessors ($N = 1$).

Students' experiences of peer assessment

Altogether 71 students completed the questionnaire, which surveyed the students' experiences of peer assessment. Most of the students considered peer assessment a good assessment practice: 64 students (90%) considered it as good or excellent, whereas only 4 students experienced it as quite or really bad, and 3 students scored "I cannot say". Most of the respondents also commented on their peer assessment experience ($N = 66$). The questionnaire results indicate that the students' experiences of peer assessment were generally positive. Based on students' descriptions, we divided the gategoris in two groups: one group of positive experiences, and the other group of negative experiences. Accordingly, altogether 39 students stated that this assessment method *supported their learning*; the following quotes describe statements in this category:

"it made you learn the subject matter more deeply"

"it was a learning situation, reflecting your own performance"

"it was very useful to me, because it made clear to me the parts of the exam that where I had trouble with"

"You got the chance to ask questions which I was thinking about after the exam. At the same time, I learned the things that I didn't know in the exam. It was a very good way to exploit the exam and also learn. I understood many thing only now (I had couple of 'Aha!' moments)".

Ten of the students noted that it was good to see how the exams were assessed: specially what is expected of them in the exam and what is assessed. Six of the students stated that it was useful to see other students' exam responses. Three students stated that it was good to receive feedback instead of just getting a number, and one student wrote that it was good to know right away how you did. Furthermore, much of the students' negative experiences of peer assessment dealt with time management, difficulty to assess other students' answers, and doubt about other students' ability to assess. Seven of the students stated that they did not have enough time to assess the two exams and listen to the discussion about them at the same time. Four of the students felt that the peer learning situation was poorly organised and the instructions were unclear. Five other students stated that the peer assessment took too much time and was a lot of work. In addition, five of the students found the assessment to be very difficult and felt that they themselves had insufficient expertise. Five of the students stated that they had doubts about the competence of the other students in assessing the exams. Finally, two of the students felt that peer assessment was useless, and three students felt that it was unpleasant to discover their own errors in the exam. The distribution of the categories appear in Table 3.

Discussion

The purpose of this study was to develop a peer assessment practice to a Bachelor level course in a bioscience context in which the assessment culture is highly traditional. The aim was to create a practice which is both valid in a sense that students understand the criteria behind the marks. The purpose was to examine the reliability of peer assessment by comparing grades given by the students and the teacher, and also to explore minor and major students' differences in their grading. The aim was also to explore how students experienced the peer assessment. Our results showed that peer assessment was highly successful in this context. The reliability of peer assessment is high, even in the beginners' course in this context. The correlations between the grades were high; the grades given by all three assessors showed no statistically significant differences. These findings are generally in line with those of a meta-analysis that compared peer and teachers marks (Falchikov & Goldfinch, 2000). Furthermore, the present study suggests that minor and major students are equally capable of assessing and grading the exams. Previous research has shown minor and major students can perform differently due to their different level of prior knowledge on the subject (Hailikari & Nevgi, 2010). However, in our study, no clear differences were found between minor and major students' evaluations and grades given by the teacher. The students in this study were presented with an assessment matrix containing detailed instructions on the evaluation and marking, and the teacher was present for the peer assessment and answered students' questions during the evaluation. Research has shown that peer assessment is more reliable when the grading criteria are well understood than when marking involves assessing several individual dimensions (Falchikov & Goldfinch, 2000; Lindblom-Ylänne et al., 2006). Thus, the present study suggests that minor students are as capable of assessing exams as major students when the instructions and assessment criteria are designed carefully and are visible to the students.

Students' experiences of peer assessment were quite positive. Altogether 90% of the students' experiences of peer assessment were either good or excellent. In addition, the open ended questions indicated that two thirds of the students felt that peer assessment supported their learning. Previous research suggests that learners tend to have a positive attitude towards assessment tasks and methods if assessment positively affects their learning and if they perceive assessment as fair (Segers, Dochy, & Cascallar 2003; Struyven et al., 2005). On the other hand, some other results indicate that the assessment culture greatly influences the way in which students experience peer learning (McGarr & Clifford, 2013). Thus, it could have been expected that many of the students would have experienced it more negatively. Nevertheless, our study suggests even when the assessment culture is very traditional the peer assessment can be a very positive experience. Although peer assessment in our study worked for summative assessment purposes and provided grades for the students, it also provided feedback for the students and helped them in their learning. The students reported that the peer assessment practice supported their learning about the subject, but it also supported students in learning important skills such as interaction skills and evaluating their own learning already in the beginning of studies. In addition to highlighting this support in learning, students also appreciated the opportunity to receive feedback on their learning and to be able to understand the criteria which served as the basis of the assessment.

Students' negative experiences focused on the organisation of and general difficulties in assessment. Some students felt that they did not have enough time to assess the exams. Others mentioned experiencing difficulty assessing their peers. One reason for these difficulties could be that these students most likely have not participated in such peer assessment before, because it is new in this faculty. In bioscience studies, the exams often measure how well the student can memorise or repeat factual knowledge (Halinen et al., 2013), especially in the beginning of studies. For this reason, students may feel that they need more time to evaluate the exams and feel uncertain about their own ability to do so. In the beginning of their university studies, students are often novices in their self-evaluation skills or critical evaluations (Smyth, 2004),

Table 3
Categories describing students' positive and negative experiences of peer assessment.

Positive experiences	*N*	Negative experiences	*N*
Supported learning	39	Not enough time	7
Clear what has been assessed	10	Took too much time	5
Useful to see others exams	6	Difficult to assess others	5
Good feedback	3	Doubt about other students' assessment skills	5
Direct knowledge about success in the exam	1	Poorly organised	4
		Unpleasant to see one's own errors	3
		Useless	2

which are among the target competences in the intended learning outcomes of academic education (Tynjälä et al., 2006). Nevertheless, the present study showed that most of the students had a very positive peer assessment experience. In this sense, our results contradict those of McGarr and Clifford (2013), whose study found that with new students, the educational value of peer learning is unclear. Our study shows that major as well as minor students in the beginning of their studies can benefit from peer assessment when students are supported in their assessment process. A study in the Faculty of Biological and Environmental Sciences has found that deep-level learning and understanding already in the beginning of studies predicts success in studies in later stages in studies (Asikainen, 2014). We therefore suggest that it is important to integrate such practices into the teaching and assessment settings so that students can practice their evaluation skills from the very beginning of the studies.

In short, the present study shows that implementing a peer assessment practice to a lecture course at the beginning of studies can lead to positive results. The present study suggests that peer assessment in grading final exams can be quite reliable in the field of biosciences already at the beginning of studies. In addition, the present study suggests that students are equally capable of assessing the exams despite having different majors and levels of achievement in the exam when the assessment process is appropriately supported. Furthermore, the present study suggests that the peer assessment of final exams is a great way to provide feedback to students and to support their learning. Students' experiences of peer assessment were very positive and emphasised its support for their learning. The high degree of agreement between the teacher's and students' marks may partly indicate that the exam mostly valued descriptive, factual knowledge. However, this emphasis on factual knowledge probably applies mostly to introductory courses in university studies, in general, or at least in science disciplines. Accordingly, our positive findings on the implementation of peer assessment are transferable not only to bioscience courses, but also to courses in other disciplines in higher education at a broader level.

Implications for practice

Along with previous research (Davey, 2011), we suggest that implementing peer assessment will help teacher more efficiently allocate their time when providing feedback and supporting students' deep learning in a large class context: meanwhile peer assessment serves summative purpose as well. Our results indicate that peer assessment could be used with high reliability also in a culture where assessment practices are traditional. Assessing a bachelor-level course in the biosciences presumably requires more quantitative than qualitative knowledge, so fixed sets of criteria can easily be designed and shared with assessors. We suggest that model answers which encourage the students to repeat their teachers' words in examinations are poorly suited to the demands of university education. Rather that providing exact answers, teachers would prefer to explain the criteria for assessment with content and other demands, such as academic writing style, in regard to the students' grading. On the contrary, peer assessment can serve as a way to foster students' evaluation skills, which is one of the main competencies required in academic expertise.

We recommend implementing peer assessment practices in large classes so that both teachers and students can benefit from them. Our results evidently confirm the effectiveness of peer assessment as an effective way to learn (Davey, 2011; Davey & Palmer, 2012) and offer a tool for increasing resources to provide feedback on student learning in a meaningful way. Our findings showed that peer assessment helped the students to better understand the course content. Thus, by marking other students' exams and through class discussion of the marks and the assessment criteria, the students were able to understand and achieve the intended learning outcomes. The activities implemented in peer assessment also clarified to the students why they received the mark they did. Thus, the students were able to receive feedback from their learning during the study process, although the teacher himself could not provide such feedback individually to each student in large class settings. However, our findings suggest that, when using peer assessment, the exams with substantially different marks given by the assessors require reassessment by the teacher to ensure reliability. Finally, out study shows that long-term pedagogical training is not the only way to develop the university teaching and learning, but a minor pedagogical support for a teacher may sometimes be enough to help students in their learning even in a context which has a traditional assessment culture.

Appendix A.

Appendix A1. The exam

Welcome to the gene technology exam!

The exam comprises three compulsory essay questions. The first question measures how well you can identify and define concepts. In the second question, the purpose is to measure your knowledge on the concepts in a deeper level by measuring how well you can compare different concepts. The third question is the most demanding which requires also application of concepts to problem solving. Remember to define the key concepts in all the essays. The essays should be clear and properly outlined. This is also taken into account in the assessment of the exam. After the exam, you will get an assessment matrix with you.

Good luck!

1. (a) In gene technology recombinant-DNA molecules are constructed by adding desired DNA-fragments into different types of vectors. Describe three different vector molecules and give examples of situations when you would use these in construction of recombinant-DNA. (4p).

(b) Probes are used in gene technology for various purposes. Describe what the term "probe" means and what are the molecular bases of its function. In addition, describe how and from what material a probe can be made and give two examples of the use of probes (6p).

2. Genes are often isolated from libraries and for this purpose two different types of libraries can be constructed; genomic libraries and cDNA libraries. Define shortly what is a genomic and a cDNA library and compare the information content in these libraries. In addition, discuss what advantages and disadvantages there is when using the libraries in isolation of genes for different purposes. (10p).

3. You are working in a clinical laboratory and your task is to use PCR technique to determine whether a mutation, predisposing its carriers to a particular genetic disease, is present in the DNA in the blood samples you are studying. The mutation is in the gene *gt15* and is caused by a single nucleotide change in the sequence, so that a C in a particular location of the sequence in the normal allele of the gene is converted to an A in the mutant allele.

Below is shown part of the sequence of the gene (500 nucleotides of the sequence is not given) and the nucleotide, which in the normal allele is C and in the mutant allele A is marked. Define shortly the principle of PCR. Describe and motivate the steps that you will use when by using PCR you determine which of the alleles the blood donors are carrying. Design also primers that you will use in your analysis (10p). *gt15* sequence (partial) CCGATGAGTGATGTGTAAACCGTCGGATAGTCGTGTCACGTGTG T**C**ACACGCAGCTGCGCGCGCGCATCGACGCGCGG—500nucleotides—CGA GCTGGACCGGCTTGAGCGTGAGCCCGGAGCTTGTGCGCGCAGTCGAC.

Appendix B

Appendix A2. The assessment matrix.

Gene technology
Number of the exam_____
Student number of the assessor_____
Grade_____
Assessment matrix

Grade/points	0 <14 Points	1 14–16 Points	2 17–19 Points	3 20–22 Points	4 23–26 Points	5 27–30 Points
Level of knowledge	Fail	Passable	Satisfactory	Good	Creditable	Excellent

	Level 1 Recognising recalling	Level 3 Understanding Comparing	Level 5 Application of knowledge	Academic writing
Exercise 1 10 points	6 Points	3 Points		1 Point
Exercise 2 10 points	3 Points	5 Points	2 Points	1 Point
Exercise 3 10 points	3 Points	3 Points	3 Points	1 Point

Justifications/comments:

Task 1. (a) In gene technology recombinant-DNA molecules are constructed by adding desired DNA-fragments into different types of vectors. Describe three different vector molecules and give examples of situations when you would use these in construction of recombinant-DNA.

Points	Assessment: mark in the exam paper which facts you cannot find in the paper. Write also your comments. Mark the total points at the bottom and also if you decide to give 1 point of scientific writing
1a Recognising Recalling 4 Points	Describing 3 different vector molecules (1 point) giving an example of their use (3 points) Examples = plasmids, phages, cosmids Plasmids are circular DNA molecules that are capable of autonomous replication in a host cell. Phages (lambda and M13) are bacterial viruses and cosmids are plasmids, containing cos-sites from phage lambda Plasmids are commonly used when DNA is multiplied in bacterial cells for further use. It is the most common vector molecule/type (1p). Phage lambda is used to generate cDNA- and genomic libraries. M13 phage is used when there is a need to get the cloned DNA in single stranded form. (1p). Cosmids are used to generate genomic libraries (1p)

(b) Probes are used in gene technology for various purposes. Describe what the term "probe" means and what are the molecular bases of its function. In addition, describe how and from what material a probe can be made and give two examples of the use of probes.

Points	Assessment: mark in the exam paper which facts you cannot find in the paper. Write also your comments. Mark the total points at the bottom and also if you decide to give 1 point of scientific writing.
1b 5 Points Recalling remembering Understanding	Explaining the meaning of probe (1point) Examples of situations that they are used (2 points) A probe is a DNA (or sometimes RNA) molecule, which is labelled by attaching to it a detectable signal and which is used to detect corresponding DNA/RNA sequences (1p) A labelled probe can be used to isolate corresponding genes from cDNA or genomic libraries. Probes can also be used to study the structure or copy number of a gene by using Southern blot or estimate the amount of the corresponding mRNA in the cells by utilizing northern blot (1p) Molecular bases of the function of a probe (2p), how and from what material a probe can be prepared (1p) The function of a probe is based on its ability to base pair with its target sequence, thus enabling the detection of the target (2p) To generate a probe a DNA-sequence corresponding to the studied gene is isolated. The probe is then prepared by DNA synthesis adding labelled nucleotides to the reaction (1p)

Task 1: points in total_____

Comments:

Task 2. Genes are often isolated from libraries and for this purpose two different types of libraries can be constructed; genomic libraries and cDNA libraries.

(a) Define shortly what is a genomic and a cDNA library.

(b) Compare the information content in these libraries.

(c) In addition, discuss what advantages and disadvantages there is when using the libraries in isolation of genes for different purposes (10p).

	Assessment: mark in the exam paper which facts you cannot find in the paper. Write also your comments. Mark the total points at the bottom and also if you decide to give 1 point of scientific writing.
2a	Explaining what is a genomic library 1, 5p
3 Points	A genomic library is a collection of genes and intergenic regions from a particular organism. The library contains all the genetic
Recalling remembering	material (the genome) of the organism and can be used to isolate and amplify genes or part of the genes
4 Points	Explaining what is a cDNA library 1, 5p
Understanding/	A cDNA library includes DNA copies of mRNA molecules present in the cells at the time when the mRNA was isolated.
comparing	A library is generated by reverse transcription of the mRNAs followed by cloning of the cDNAs into a lambda vector
2b	Comparing the information content in these libraries 2p
2 Points	Genomic library is composed of all the genes in an organism and contains also the promoter regions and introns of the genes.
Oma pohdinta	Because cDNA libraries are generated by conversion of the isolated mRNAs into cDNAs they only contain the parts of the genes that are present in mature mRNAs, therefore lacking the promoter regions and introns. In addition a cDNA library contains only part of the genes in an organism, harboring only copies of those genes that were expressed in the cells from which the mRNA was isolated. (Whereas a genomic library contains all of the genes, not only those that are expressed)
	Comparison of use 2p
	Genomic library is used in situations where all the possible information regarding a gene is wanted. For example when studying the whole genetic makeup of an organism or when the structure of a particular gene is analyzed (promoters/introns etc.). cDNA library is used when the interest is in gene expression or in the proteins encoded by the genes
	Reflecting advantages and disadvantages 2p
	Genes isolated from genomic libraries are structurally similar to the genes in the genome, allowing structural studies of the genes only for clones isolated from these libraries. On the other hand, due to the large amount of information present in genomic libraries, the handling of the libraries (especially screening of them) is more labor intense than handling of cDNA libraries. cDNA-libraries never contain information for all of the genes in an organism and their information is only from the parts of the genes that are present in the encoded mRNAs. The incompleteness of cDNA libraries can also be seen as an advantage, because a cDNA library generated from specific cells is more compact than a genomic library but harbor all the genes of interest (2p)

Task 2 points in total_____
Comments:

Task 3. You are working in a clinical laboratory and your task is to use PCR technique to determine whether a mutation, predisposing its carriers to a particular genetic disease, is present in the DNA in the blood samples you are studying. The mutation is in the gene *gt15* and is caused by a single nucleotide change in the sequence, so that a C in a particular location of the sequence in the normal allele of the gene is converted to an A in the mutant allele.

Below is shown part of the sequence of the gene (500 nucleotides of the sequence is not given) and the nucleotide, which in the normal allele is C and in the mutant allele A is marked. Define shortly the principle of PCR. Describe and motivate the steps that you will use when by using PCR you determine which of the alleles the blood donors are carrying. Design also primers that you will use in your analysis (10p)

gt15 sequence (partial) CCGATGAGTGATGTGTAAACCGTCGGATAGTCGTGTCACGTGTGT**C**ACACGCAGCTGCGCGCGCGCATCGACGCGCGG—500nucleotides—CGAGCTGGACCGGCTTGAGCGTGAGCCCGGAGCTTGTGCGCGCAGTCGAC.

Points	Assessment: mark in the exam paper which facts you cannot find in the paper. Write also your comments. Mark the total points at the bottom and also if you decide to give 1 point of scientific writing.
3 Points	Definitions of PCR technique 3 points
Recalling remembering	PCR (polymerase chain reaction) is a method that can be used to amplify a defined region of a DNA molecule. PCR-technique is
4 Points	based on cyclic DNA synthesis by using a heat stable DNA polymerase and two primes that delineate the amplified region. PCR cycle
Understanding/comparing	contains three steps, which are regulated by temperature alterations. The cycle: (1) DNA:*n* denaturation (94 °C) (2) annealing of
3 Points	primers (~50 °C) (3) DNA synthesis (72 °C)
Applying knowledge	PCR reaction is performed in a machine, where the temperatures, reaction times and the amount of cycles are programmed
1 Point	Explaining the bases of allele determination (motivation of different steps) 4 points
	DNA isolation and PCR reaction by using two pairs of primers (allele specific + conserved) (1p)
	The use of allele specific primers leads to DNA amplification only in the reaction where the nucleotide in the DNA is the same as in the primer (1p)
	The reason for differential amplification is the fact that in DNA synthesis the last nucleotide of the primer needs to be base paired with the template for the synthesis to start (1p)
	Designing primesrs 3points Allele specific primers (1p) 5′CCGTCGGATAGTCGTGTC-3′ ja 5′CCGTCGGATAGTCGTGTA-3′
	Primers are identical, except the last nucleotide, which is allele specific
	Conserved primer ~18 nucleotides from anywhere after the 500 nt:n of unknown sequence (1p). The primer needs to have reverse complement sequence of the provided sequence (2p)
	The primer can be e.g., 5′-CCGGGCTCACGCTCAAGC-3′
	Scientific writing

Task 3: points in total _____Comments:

References

Anderson, L. W. (2001). In D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. Raths, & M. C. Wittrock (Eds.), *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives (Complete edition)*. New York, NY: Longman.

Asikainen, H. (2014). Successful learning and studying in biosciences. Exploring how students' conceptions of learning, approaches to learning, motivation and their experiences of the teaching-learning environment are related to study success. In *Doctoral dissertation* Retrieved from e-thesis database ⟨http://urn.fi/URN:ISBN:978-952-10-9367-8⟩.

Asikainen, H., Parpala, A., Virtanen, V., & Lindblom-Ylänne, S. (2013). The relationship between student learning process, study success and the nature of assessment. A qualitative study. *Studies in Educational Evaluation, 39*(4), 211–217.

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment and Evaluation in Higher Education, 27*(5), 427–441.

Black, P., Harrison, C., Lee, C., Marshall, B., & William, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 9–21.

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment and Evaluation in Higher Education, 4*(31), 399–413.

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

Bryan, C., & Clegg, K. (2006). *Innovative assessment in higher education*. New York, NY, London: Routledge.

Davey, K. (2011). Student peer assessment: Research findings from a case study in a master of chemical engineering coursework-program. *Education for Chemical Engineers, 6*(4), 122–131.

Davey, K., & Palmer, E. (2012). Student peer assessment: A research study in a level III core course of the bachelor chemical engineering program. *Education for Chemical Engineers, 7*, 85–104.

Elliot, N., & Higgins, A. (2005). Self and peer assessment—dies it make a difference to student group work? *Nurse Education in Practice, 5*, 40–48.

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.

Gouli, E. (2008). Supporting self-, peer-, and collaborative-assessment in E-learning: The case of the PEer and collaborative Assessment environment (PECASSE). *Journal of Interactive Learning Research, 4*(19), 615–647.

Hailikari, T., & Nevgi, A. (2010). How to diagnose at-risk students in chemistry: The case of prior knowledge assessment. *International Journal of Science Education, 32*(15), 2079–2095.

Hailikari, T., Nevgi, A., & Lindblom-Ylänne, S. (2007). Exploring alternative ways of assessing prior knowledge, its components and their relation to student achievement: A mathematics based case study. *Studies in Educational Evaluation, 33*(3–4), 320–337.

Halinen, K., Ruohoniemi, M., Katajavuori, N., & Virtanen, V. (2013). Life science teachers' discourse on assessment: A valuable insight into the variable conceptions of assessment in higher education. *Journal of Biological Education, 48*(1), 16–22.

Hinett, K. (1999). Students' get a fair chance to make a mark. In *Times higher education supplement* ⟨http://www.timeshighereducation.co.uk/news/students-get-a-fair-chance-to-make-a-.mark/148596.article⟩.

Lindblom-Ylänne, S., Pihlajamäki, H., & Kotkas, T. (2006). Self-, peer-and teacher assessment of student essays. *Active Learning in Higher Education, 7*(1), 51–62.

Lueddeke, G. (2003). Professionalising teaching practice in higher education: A study of disciplinary variation and 'teacher-scholarship'. *Studies in Higher Education, 28*(2), 213–228.

McGarr, O., & Clifford, A. M. (2013). 'Just enough to make you take it seriously': Exploring students' attitudes towards peer assessment. *Higher Education, 65*, 677–693.

Nicol, D., & Macfarlane-Dick, D. (2006). Formative assessment and self regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218.

Nulty, D. N. (2011). Peer and self-assessment in the first year of university. *Assessment and Evaluation in Higher Education, 36*(5), 493–507.

O'Moore, L. M., & Baldock, T. E. (2007). Peer assessment learning sessions (PALS): An innovative feedback technique for large engineering classes. *European Journal of Engineering Education, 32*(1), 43–55.

Ploegh, K., Tillema, H., & Segers, M. S. R. (2009). In search of quality criteria in peer assessment practices. *Studies in Educational Evaluation, 35*, 102–109.

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

Prins, F. J., Slujsmans, D. M., Kirchner, P. A., & Strijbos, J.-W. (2005). Formative peer assessment in a CSCL environment: A case study. *Assessment and Evaluation in Higher Education, 30*(4), 417–444.

Räisänen, M., Tuononen, T., & Postareff, L. (2012). What do course grades tell us about the quality of learning outcomes in biosciences? *Paper presented in Sig Higher Education Conference*.

Sadler, R. (2009). Grade integrity and the representation of academic achievement. *Studies in Higher Education, 34*, 807–826.

Segers, M., & Dochy, F. (2001). New assessment forms in problem-based learning: The value-added of the students' perspective. *Studies in Higher Education, 26*(3), 327–343.

Segers, M., Dochy, F., & Cascallar, E. (2003). *Optimising new modes of assessment: In search of qualities and standards*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Sluijsmans, D., & Prins, F. (2006). A conceptual framework for integrating peer assessment in teacher education. *Studies in Educational Evaluation, 32*(1), 6–22.

Smyth, K. (2004). The benefits of students learning about critical evaluation rather than being ummatively judged. *Assessment and Evaluation in Higher Education, 29*(3), 370–378.

Somervell, H. (1993). Issues in assessment, enterprise and higher education: The case for self-peer and collaborative assessment. *Assessment and Evaluation in Higher Education, 18*(3), 221–233.

Stiggins, R. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 38*(10), 758–765.

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment and Evaluation in Higher Education, 30*, 325–341.

Tynjälä, P., Slotte, V., Nieminen, J., Lonka, K., & Olkinuora, E. (2006). From university to working life: Graduates' workplace skills in practice. In P. Tynjälä, J. Välimaa, & G. Boulton-Lewis (Eds.), *Higher education and working life—Collaborations, confrontations and challenges* (pp. 73–88). Amsterdam: Elsevier.

Welsh, M. M. (2007). Engaging with peer assessment in post-registration nurse education. *Nurse Education in Practice, 7*, 75–81.

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education, 45*(4), 477–501.

Henna Asikainen, PhD, Post-doctoral researcher.

Viivi Virtanen, PhD, Pedagogical university lecturer.

Liisa Postareff, PhD, Univercity lecturer.

Pekka Heino, PhD, University lecturer.