Third Time is a Charm – Determining the Required Number of Assessors when Using Peer Assessment in Large-Scale Lectures

Completed Research Paper

Katja Lehmann[1]

Matthias Söllner[1,2]

Ivo Blohm[2]

Jan Marco Leimeister[1,2]

[1] University of Kassel, Research Center for IS Design (ITeG), Information Systems, Pfannkuchstr. 1, D-34121 Kassel, Germany
[soellner;leimeister]@uni-kassel.de
[lehmann]@wi-kassel.de

[2] University of St.Gallen, Institute of Information Management (IWI-HSG), Mueller-Friedberg-Str. 8, CH-9000 St.Gallen, Switzerland
[matthias.soellner;ivo.blohm;janmarco.leimeister]@unisg.ch

Abstract

Although peer assessment is a widely used didactical method in higher education, little is known about a) how many peer assessors are required to receive a stable assessment on a peer's solution and b) how valid this assessment is compared to an expert assessment. To fill these gaps, we conducted a peer assessment in a large-scale lecture. Overall, 136 students participated in the peer assessment. Each student had to complete an assignment, which was then anonymously evaluated by five randomly selected peers, and three independent experts. We applied a bootstrap-based Monte Carlo simulation based on our data. The results show that a) three peer assessors are sufficient for a stable assessment, and b) the peer assessors are less critical compared to experts. We thus contribute to literature by providing insights on how many peer assessors are required when applying peer assessment, and how comparable peer assessment is with expert assessment.

Keywords: Complex assignments, Expert assessment, Large-scale lectures, Monte Carlo simulation, Peer assessment.

Introduction

Large-scale lectures are still common default at universities worldwide, especially in basic lectures at an undergraduate level (Fortes and Tchantchane 2010). An auditorium with 100 or more students is not an uncommon situation leading to a lecturer-centered teaching style. Those lectures lack individual feedback within the teaching-learning process (Bligh 2000; Lehmann and Söllner 2014). Furthermore, the imparting and verification of knowledge aims primarily at factual knowledge (Hagstrom 2006) addressing the low cognitive levels of educational objectives (namely remembering, understanding, and applying with regard to Anderson et al. (2001)). Focusing on factual knowledge regarding the low cognitive levels of educational objectives is not sufficient with regard to the further study process as well as considering the complexity of future tasks as employees in the business world (Knight and Wood 2005). Therefore, the high cognitive levels of educational objectives distinguished by Anderson et al. (2001), namely analyzing, evaluating, and creating, have to be taken into account for the imparting and verification of knowledge. This is relevant to comprehend the learning contents in all their complexity (Knight and Wood 2005).

However, the verification of knowledge at the high cognitive levels of educational objectives distinguished by Anderson et al. (2001) is characterized by complex assignments (e.g., extensive free-text assignments, writing statements, and essays). Consequently, the verification of those assignments is time and resource consuming and, hence, impossible to use within the teaching-learning process in a large-scale lecture. This, in turn, calls for the lecturers to offer complimentary didactical methods centering on the students and supporting them in their teaching-learning process by providing feedback on the individual learning progress. One promising approach to enhance interaction, feedback and to assess high cognitive levels of educational objectives without massively increasing the workload of lecturers is the use of peer assessment (Strijbos et al. 2009). By using peer assessment, students evaluate the quality of another student's performance during the teaching-learning process (Tahir 2012) according to specifically defined criteria (Boud and Falchikov 2007). Consequently, peer assessment makes imparting and verification of knowledge concerning the high cognitive levels of educational objectives in large-scale lectures feasible and manageable since correcting the assignments is no longer the task of the lecturer. Using peer assessment, the students are encouraged to work independently and to develop skills in high cognitive areas (Bostock 2004), thereby reducing the workload of the lecturer (Sadler and Good 2006). Students, in comparison to the lecturer, never assess all their peers' assignments, but only a certain number of them. Although several studies examined the method of peer assessment in comparison to the assessment carried out by experts in various settings (Cheng and Warren 1999; Hovardas et al. 2014), so far no study compared peer and expert assessment in the context of the verification of the high cognitive levels of educational objectives. Moreover, the results comparing peer and expert assessment differ widely depending on the peers' experiences of evaluating others, the educational levels of the students, the used assessment form or the different assessment procedures employed (Chang et al. 2012). Consequently, this leads to our first research question which is as follows:

> *RQ1: How comparable are the assessments by peer assessors to the assessments carried out by the expert assessors?*

Besides the benefits the peer assessment has, it brings additional effort for the students due to the fact that they are highly engaged when participating in the peer assessment. However, it is not the intention to overload the students by putting too much pressure on them in providing feedback to a multitude number of students, which can result equally in frustration, demotivation and decreasing willingness to participate. When students think that they have to do too much or irrelevant learning activities it increases the cognitive load (Eveland and Sharon 2000) and split student's attention (Harter 1986). While conducting the peer assessment makes it necessary to engage a great number of students, the students might feel overload when they think they provide feedback on the solutions of too many other peers. Therefore, it is necessary to investigate how many peer assessors are required to receive a stable assessment on another peer's solution. This provides the justification for the study at hand. The challenge is to investigate the tipping point that indicates the number of peer assessors who are needed for a valuable contribution in the design of the peer assessment. The risk is that by using an even higher number of peers might decrease student motivation and their willingness to participate in the peer assessment. On the other hand, using fewer peers could decrease the quality of the aggregated peer assessment because the feedback may contain less diverse information. This leads to the second research question for this study:

> *RQ2: How many peer assessors are required to receive a stable assessment on a peer's solution?*

We use data to simulate the number of peer assessors required to receive stable aggregated assessments. To assess the tipping point that indicates the required number of peer assessors, we performed a Monte Carlo approximation to the bootstrap estimate to determine how many peer assessors per assessment object are required to receive a stable assessment on a peer's solution. To the best of our knowledge, no studies investigate how many peer assessors are required for a stable assessment on a peer's solution concerning the use of complex free-text assignments addressing the high cognitive levels of educational objectives. Hence, the result of this study closes a research gap and at the same time provides implication for the design of a peer assessment with an optimal use of resources on the part of the students.

For answering our research questions, we conducted a peer assessment in an introductory large-scale lecture on information systems for undergraduate business administration students. To generate enough data, we used a 1:5 peer assessment, meaning that each solution was evaluated by five peers. Overall, 136 students participated in the peer assessment. Furthermore, each solution was assessed independently by three experts. The lecturer assistant and two tutors, who supported the lecture for several semesters, served as expert assessors. Therefore, our data sample comprises five peer assessments and three expert assessments for each student solutions. The peer assessment was used in a quasi-experimental setup during the teaching-learning process with a four-week duration and voluntary participation. The system of use was the learning management system (LMS) Moodle with the workshop module.

The remainder of our paper is organized as follows: first, we outline the existing literature regarding the use of peer assessment in higher education and refer to the role of the students as assessor and as assessee, and the resulting benefits. Moreover, we present related work on research studies comparing peer and expert assessment. Afterwards, we describe the methodology of our study, and present the results. Finally, we discuss our results and refer to the implications for literature on peer assessment.

Theoretical Background

Peer Assessment in Higher Education

Peer assessment allows for individual feedback on the learning success as well as corresponding interventions even in groups with a higher number of students (Falchikov and Goldfinch 2000). The assessment of the value or quality of another student's performance is realized according to specifically defined criteria for quantitative feedback (Boud and Falchikov 2007), or comments for qualitative feedback, or both together (Hsia et al. 2016). When students are supposed to provide feedback to other students, pre-defined feedback criteria are essential in order to support the feedback provider and to ensure that valuable and constructive feedback is given (Falchikov and Goldfinch 2000). Students are getting involved in the teaching-learning process by providing and receiving feedback to and from peers who work on the same assignment to help each other to enhance the individual learning performance (Tahir 2012). We follow these descriptions and use the term of peer assessment to describe students of a peer group mutually evaluating each other's performances according to defined criteria, while formulating an overall written feedback including strengths, weaknesses, and suggestions.

Peer assessment is not only applied in educational settings. Pair programming is a common method in computer science, where software developers control each other's work and point out mistakes or complicated designs (Umar and Hui 2012). Scientists apply peer review to assess other scientists' conference papers in order to ensure quality (Falchikov and Goldfinch 2000).

However, in higher education the application of peer assessment generates especially the following benefits as opposed to an evaluation conducted solely by the lecturer:

- Increased Efficiency: Lecturers save valuable time if students provide each other with feedback and evaluate each other's academic performance (Sadler and Good 2006).
- Pedagogically: The evaluation of responses regarding correctness gives the student a deeper understanding of the learning contents. By reading solutions of others, they can expand their own knowledge and develop new ideas by evaluating other points of view (Sadler and Good 2006).
- Metacognitively: Students will develop an awareness for their own strengths and weaknesses (Tahir 2012) and will be able to compare and evaluate their own performance with that of their peers, at least to a certain extent (Darling-Hammond et al. 1995). Therefore, students may reach a more accurate self assessment when participating in a peer assessment (Topping et al. 2000). In addition, students train their ability to think critically as well as their evaluation and reflection skills (Jaillet 2009).
- Affectively: Students perceive qualitative feedback from the peer group as more valuable than a lecturer's grade (Sadler and Good 2006).

Therefore, the application of peer assessment does not only relieve the lecturer, but turns students into experts themselves.

The Student as Assessor and as Assessee

Conducting peer assessment each student adopts both roles – the role as assessor and the role as assessee. The assessment of the performance of other students provide valuable insights for the own learning and allows a useful learning experience. For supporting students in their role as assessor, several requirements are necessary: criteria need to be defined which help the assessors in conducting the assessment with constructive feedback. The assessors need criteria to measure and assess the performance of the assessee. Typically, criteria are provided in the form of a scale (Hafner and Hafner 2003) that help the peer assessors to decide whether the peer's solution is a good or weak one (Tseng and Tsai 2007). Depending on the type the peer assessment is conducted the assessors need to have writing or communicating skills. Each assessor is responsible for making critical judgments in order to assess and to analyze the performance of a peer, by applying the assessment criteria (Topping et al. 2000). Therefore, the assessor has to be able to communicate and to justify the own assessment.

Acting as peer assessor involves tasks in reviewing, summarizing, clarifying, providing feedback, recognizing misconceived knowledge, identifying missing knowledge, and diagnosing deviations from the ideal (Van Lehn et al. 1995). Therefore, peer assessment demands highly cognitive activities from the students. Operating as peer assessor is associated with various potential benefits for the own learning and the training of essential personal skills. The assessors train their assessment and judgement skills when thinking critically and evaluating the solution of others (Leijen et al. 2009). By having the opportunity to assess the solution of others, the assessor gets an awareness for the own strengths and weaknesses (Tahir 2012), which increases the skill for self-reflection. Being confronted with other points of view might lead to own new ideas and knowledge extension (Chen 2010; Sadler and Good 2006). When contributing feedback, it is highly relevant to provide as much information as possible to the peers and to clearly outline the next steps to further improve the solution (Tsivitanidou et al. 2011). Another benefit refers to the opportunity to participate in important cognitive activities, such as critical thinking (e.g., to come to a decision which piece of work is a good or weak one), planning, monitoring, and regulation (Hovardas et al. 2014). Another benefit refers to receiving feedback regarding understanding of what the received assessment is about and to better estimate its relevance and purpose (Brindley and Scoffield 1998). Moreover, by participating in the peer assessment students will receive a better understanding regarding the requirements for specific assignments and they will get an awareness of what is required to attain a particular performance level (Hovardas et al. 2014).

Many research studies that focus on peer assessment pay attention on psychometric characteristics and rather draw a positive picture of peer feedback as assessment devices. Feedback from more than one person allows for the aggregation of evaluations by virtue of their number in contrast to the assessment from a single assessors (e.g., the professor) (Brutus and Donia 2010). The pooling of assessments from several peers refers to an increase in reliability and a partial removal of biases that are connected with a single assessor (Conway and Huffcutt 1997; Greguras and Robie 1998). Peer assessment also shows validity according to several research studies (Gardner et al. 1998; Jaillet 2009; Schumacher et al. 1992). Regarding the evidence of reliability and validity, peer assessment indicates also a significant impact on individual learning processes (Brutus and Donia 2010). Hence, integrating peer assessments in the teaching-learning process as a didactical method can have a positive influence on both assessor and assessee.

When a student changes the role from an assessor to an assessee the required capabilities are also changing. The student has to assess and to review the received feedback. A critical review has to carry out which part of the feedback is valuable and which part contains any misconceptions or errors (Tsivitanidou et al. 2011). Otherwise, the students will not trust the peers comments (Söllner et al. 2016), and might reject them even though they were correct. Afterwards, a decision has to be made which part of the own solution has to be changed according to the received feedback. The proposed changes has to revise to further improve the own solution (Hovardas et al. 2014; Tsivitanidou et al. 2011). When the received feedback clearly evinces the errors and provides advice for improvement which are feasible to implement then the feedback is constructive and helpful (Hovardas et al. 2014; Topping 1998). However, the received feedback can be experienced as inadequate for various reasons. One reason could be when the assessor themselves is not self-confident regarding the topic which is dealt with in the assignment of the peer assessment. Another reason might be when the assessor is self-conscious regarding the own assessment skills, because it is the first feedback provision (Gielen et al. 2010).

The Decision Process for Feedback Provision

The assessment process requires an investigation of the peer's solution, to evaluate it in relation to clever questions at a macro and micro level (Graesser et al. 1995). Providing feedback on a peer's solution with criteria and rating scales the students runs a complex cognitive process that is similar to responding to a survey. When answering the questions in a survey, first the respondents have to understand the item, make an individual judgment regarding that item and afterwards, select one of the provided rating scales to express their own judgment on a specific item. However, this process can be transferred to what the students have to do in providing quantitative feedback on another peer's solution. Tourangeau et al. (2000) provide a decision process consisting of four steps and it can be assumed that the peer assessors use this decision process when providing feedback. The first step is the so-called comprehension, meaning to understand the item, to become aware of the meaning, and inferring the item's point (Tourangeau et al. 2000). The second step encompasses the retrieval of relevant information by creating a retrieval strategy, using signals for information recall as well as to remember generic and specific memories and filling in missing details (Tourangeau et al. 2000). The third step in the decision process refers to the judgement that has to be done regarding the items. Specifically, the retrieved information are judged concerning relevance and completeness in the context of the overall judgment (Biemer and Lyberg 2003; Tourangeau et al. 2000). During the decision process the initial judgement is changing with regard to the further judgement of the following items (Riedl et al. 2013). In the last step, the so-called reporting and response selection, the respondents indicate their judgement by choosing a specific rating scale. After this process of encoding, the decision process ends. However, the responses may be biased in regard to consistency with previous responses, social acceptability or other influencing factors (Biemer and Lyberg 2003). The decision process is not set fix and allows switching to subsequent stages during the rating (Biemer and Lyberg 2003; Tourangeau et al. 2000).

Related Work on Peer Assessment in Comparison to Expert Assessment

Several studies show that peers are indeed able to provide valuable feedback (Dochy et al. 1999; Falchikov and Goldfinch 2000). There are several research studies considering peer assessment in higher education, usually focusing on scores and grades given by the peers rather than on focusing on open-ended questions. However, research studies observed different results regarding the comparison of peer assessment and the assessment carried out by the lecturers, who are the experts.

In the literature review conducted by Topping (1998), he identified that in 18 of 25 research studies a high reliability was reported in comparing marks or grades provided by the lecturer and the peers. The investigated research studies measured the agreement between the lecturer and the peers with correlation coefficients, percentage agreement, or with a measurement of central tendency and variance, sometimes with statistical significance (Topping 1998). Hughes and Large (1993) noted a high correlation between the assessment carried out by the peers and the assessment carried out by the staff. The authors used the peer assessment method within oral presentations and the peers had to assess their communication and presentation skills – a procedure they are familiar with. Kulkarni et al. (2013) used peer assessment in a massive online class where students had to create artifacts like paper prototypes. The authors observed the grades awarded by the peers highly correlated with the grades awarded by the lecturer.

Research studies that investigated no agreement between the peer and the lecturer assessment exist in the area of essay writing (Mowl and Pain 1995) and in group project assignments (Boydell 1994; Mathews 1994). By comparing grading results, Lin et al. (2001) indicated that the lecturer assessment is more strict in comparison to the peer assessment. The same result achieved Chang et al. (2012), who compared students in high school with their achieved portfolio grades and detected that the grades assigned by the teacher are stricter than those assigned by the peers.

However, the conditions in several of the research settings under which the peer assessment occurs differ greatly. The achieved results when comparing peer and lecturer assessment vary from one study to another depending on the peers' experiences in providing feedback, the educational levels of the students, the used assessment form or the different assessment procedures (Chang et al. 2012). Hence, a general conclusion whether peer assessments and expert assessments are always comparable is not possible since several effects influence the results. In our setting we used peer assessment for the first time. Hence, the students have never assessed their peers before. Additionally, we used a complex assignment for the peer assessment. Therefore, we refer to the result achieved by Chang et al. (2012) and assume that the students will assess their peers less strict in comparison to the experts.

Method

Participants

The sample was composed of undergraduate business administration students from a large public university. The lecture in which we used the peer assessment was an introductory large-scale lecture on information systems. The students were in their first or second year of study. The lecture is a compulsory lecture. Within this lecture it was the first time that the students had to solve a complex free-text assignment individually and it was the first time that the peer assessment was conducted.

The participation in the peer assessment was voluntary. 152 students attended the peer assessment (participation rate is of 86.21% based on 174 students who took part in the final exam). However, there were at least several students who submitted a solution but did not provide any feedback on a peer's solution. These were excluded for our data analysis which reduces the sample size. Therefore, we refer to a sample size of 136 students, which means that for each of the 136 student's solution we have five assessments of the peers. The mean age of the peer assessors was 23.05 years, and 54.41 percent were female. To ensure that self-selection bias is not a serious issue in our study, we conducted a two-way comparison of the treatment group (students that participated), and the control group (students who did not participate). The results show that both groups do not differ significantly in terms of demographics, as well as in the exam performance besides the free-text assignment we intended to improve with our assignment (please see Appendix A for further details).

Subsequently to the peer assessment three expert assessors (one female, two male) rated each of the peers' solution. We used the independent expert assessment as a baseline for comparison to assess the quality of the peer assessment. The lecture assistant and two tutors served as expert assessors. The tutors are students who support the lecture regarding accompanying tutorial groups since several semesters. Specifically, one tutor supports the lecture for four years and the other tutor since one and a half year. All three expert assessors are familiar with the learning content of the specific lecture. Additionally, all expert assessors have significant prior experience in grading students work. For each of the 136 student's solution we have three assessments by the experts.

Quasi-Experimental Task and Procedure

The peer assessment was used once throughout a four-week period during the teaching-learning process (Lehmann and Leimeister 2015; Lehmann et al. 2016). Lecture grades were unaffected by the peer assessment participation. The lecture grade for each student was fully determined by the final exam. Due to legal regulations we were not able to provide credit points within the teaching-learning process. The acquisition of credit points is possible in the final exam as summative assessment only. But, we were allowed to provide bonus points within the teaching-learning process. Therefore, participation in the peer assessment was rewarded with one bonus point extra on top of the final exam scores, which has 90 points in total.

The LMS Moodle with the workshop module was used as technical support in our lecture because it constitutes the essential eLearning infrastructure at our university. Thereby, we used an existing platform and adjusted the settings regarding our needs. A schedule set certain deadlines, with each deadline instructing the students on what to do in what time frame. Additionally, short videos explained how to use the workshop module in the LMS in order to avoid operational problems. Furthermore, the students were reminded to solve the remaining task before each deadline. Figure 1 schematically illustrates the peer assessment process.

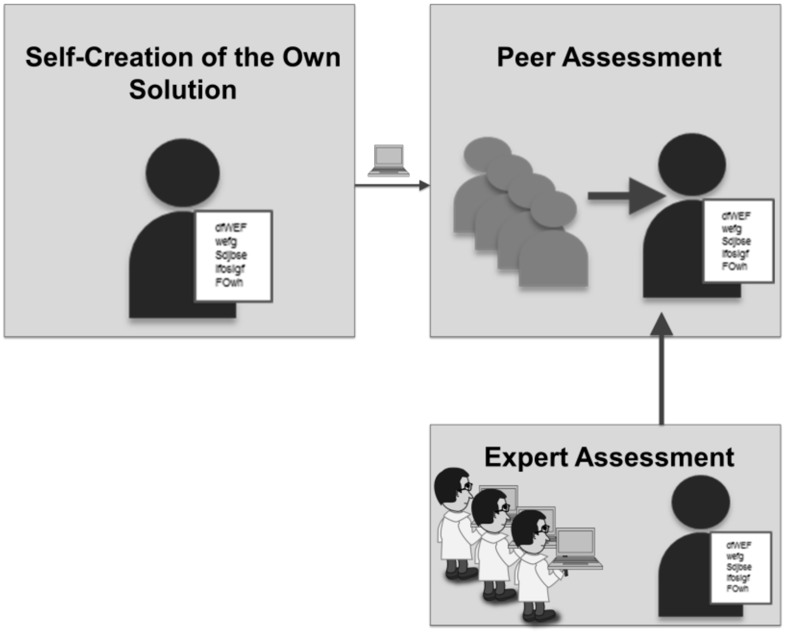

Figure 1. Schematic Process of the Procedure for Peer and Expert Assessment

Self-Creation of the Own Solution

In a first step, the students receive an extensive free-text assignment for individual work to interact with the learning content and to create an own solution. Each student had to solve the free-text assignment individually within one week before uploading it to the LMS. The free-text assignment for the peer assessment required depth of content for the solution development, the combination of learning contents, as well as formulating own arguments in the form of a statement. Hence, the free-text assignment used addressed the high cognitive levels of educational objectives.

The topic of the free-text assignment required the creation of a cost-benefit analysis for a self-chosen example of a business software delivery. The students were expected to find criteria for the cost-benefit analysis on their own, and to rate and discuss them. In addition, they were asked to discuss the usefulness of a cost-benefit analysis regarding the software delivery.

Each student had to solve the assignment without writing the own full name in the solution in order to guarantee anonymity. This was ensured by requiring each student to create an own code for their solution based on first and second letters of own mother's first name and the own street number.

All students received the same assignment. Regarding the page limit the students got a specification of four typewritten DIN A4 pages in a Word file, Arial and font size 11. Apart from answering the free-text assignment, the students were asked to work on final exam-related assignments throughout the semester.

Peer Assessment

The LMS automatically distributed the assignment to five randomly selected students. Thus, 1:5 peer assessment was applied meaning that each solution was evaluated by five different peers and each student provided feedback on solutions submitted by five fellow students. The reason for the application of the 1:5 assessment was to investigate the required number of peer assessors. The assessment was fulfilled anonymously, that means no student was aware whose solution they were assessing. This way, the feedback is more precise, valuable, and honest (Bostock 2004). Students feel less negative emotions such as peer pressure or fear of disapproval (Tahir 2012) and a possible feedback manipulation based on social relationships is avoided (Boud and Falchikov 2007). Hence, anonymous feedback allows more content-based and objective feedback. The students were guided with a tutorial how to provide feedback to their peers. The provided feedback had to be uploaded to the LMS by a certain deadline.

After receiving the peer's feedback each student had to revise their solution accordingly. Consequently, each student received feedback regarding the own performance on the specific free-text assignment and was able to align the self-assessment with the assessment of the peers.

Expert Assessment

To assess the quality of the peer assessors, we compared their assessment with three independent expert assessors. We did this intentionally to answer our research questions. The expert assessment was collected after the peer assessment. The students did not receive the feedback from the expert assessors.

All solutions of the peers were assessed by the qualified expert assessors using the consensual assessment technique (Amabile 1996). Relevant literature states that an expert jury of three to ten people is sufficient in order to obtain a reliable assessment of the results (Amabile 1996).

We decided to use three persons as expert assessors, which is in line with the lower limit of the requirement following Amabile (1996). This was the maximum number we were able to engage for the expert assessment according to our resources and who fulfilled the further requirements of Amabile (1996); in specific regarding the requirement of familiarity. The expert assessors showed high familiarity with the object of investigation. The familiarity is based on the fact that the lecturer and the two lecture assistants work for that specific lecture since several semesters and are experienced in evaluating students' performance regarding the specific learning content of the lecture. Each expert assessor worked on their own and independently of one another. They were informed not to communicate among each other regarding the evaluating of the student's solution which was ensured through separated workspaces so that each expert worked in a room on their own. Since each student used a specific code for the own solution to guarantee anonymity the expert assessors had no awareness whose solution they were assessing.

Data Collection

The assessment form contained space for a qualitative, written feedback and a rating sheet to provide quantitative feedback regarding several criteria. For our research study, only the results of the rating sheet are of interest.

The rating sheet comprised criteria concerning the author's knowledge expertise and the solution's quality. Additionally, the rating sheet captured criteria for assessing the peer's knowledge on the high cognitive levels of educational objectives. The rating sheet comprised only items according to the training objectives of the conducted peer assessment.

To assess the knowledge expertise apparent in each solution, we used a set of six items based on Braun et al. (2008). Both peer and expert assessors had to assess the student's skills regarding the learning content. Specifically, each assessor had to assess whether the student is able to provide an overview regarding definitions and concepts of the learning content, to compare differences and similarities of the learning content and to solve a typical problem related to the learning content.

The quality of the solution was measured using a set of six items adopting from Bauer et al. (2009). Whether the solution was coherent and easy to understand, whether the solution answered the question of the free-text assignment accurately and whether the argumentation was of high quality contributed to the decision about the quality of the solution.

For assessing the peer's knowledge on the high cognitive levels of educational objectives, six items were newly created. We followed the description of each cognitive level of educational objective based on Bloom and Krathwohl (1956) and Anderson et al. (2001). Therefore, each assessor had to assess whether the author of the solution is able to remember, to understand, and to apply the specific learning content of the assignment. Additionally, each solution had to be assessed according to the author's ability to analyze the learning content, to recognize relationships as well as to deduce inferences and to come to a decision.

Moreover, the assessment focused on whether the author of the solution was able to create a new solution by composing knowledge components.

All 18 items in the assessment form were measured on a 7-point Likert scale ranging from 1 = *Strongly Disagree* to 7 = *Strongly Agree*. Additionally, the students had the option to choose "no comment".

Both peer and expert assessors used the same assessment form for their assessments. In total, the 136 author's solution generated 19.584 data in total (136 solutions x 5 peer assessors x 18 items + 136 solutions x 3 expert assessors x 18 items).

Results

Measures of Reliability for Peer Assessment and Expert Assessment

In a first step, we refer to the reliability of the peer assessment as a variable regarding Falchikov et al. (2000). Therefore, we want to investigate to which extent the peer assessors and the expert assessors interrelate among each other. Since our variables are interval scaled, we examined the intraclass correlation to explore whether the peer and the expert assessors' evaluations show a resemblance. Specifically, we used the intraclass correlation coefficient (ICC) to examine the interrater reliability of how consistent the five peer assessors and the three expert assessors are among each other. The tool of analysis was SPSS 23. The ICC is based on values from -1.0 to +1.0. The reliability is considered as acceptable if it is better than .50, which shows a good performance by multiple raters (O'Neill et al. 2012).

For reliability of the peer assessors and in the formal term of Shrout and Fleiss (1979), the analysis used the ICC(1,k). This means that the raters were selected randomly and each case was not rated by each peer assessor. Each of the peers' solution was rated by five randomly selected peers. The result shows a good reliability of the peer assessors among each other (.73).

To investigate whether the expert assessors interrelate among each other, we measured the ICC(3,k), which means that all of the three expert assessors rate each case. We measured consistency by using the mean value of each case. Therefore, we used the two-way mixed average measure. Our analysis shows a good reliability of the expert assessors among each other since the ICC(3,k) has a value of .77. In addition, the requirements of Amabile (1996) for the expert jury were fulfilled.

Table 1. Intercorrelations between Peer Assessment and Expert Assessment

	1	2	3	4	5	6	7	8
1. Experts – Knowledge Expertise	1							
2. Experts – Solutions Quality	.967**	1						
3. Experts – Educational Objectives	.969**	.969**	1					
4. Experts – Overall	.988**	.988**	.989**	1				
5. Peers – Knowledge Expertise	.637**	.621**	.600**	.627**	1			
6. Peers – Solutions Quality	.642**	.626**	.618**	.632**	.901**	1		
7. Peers – Educational Objectives	.653**	.634**	.637**	.647**	.925**	.932**	1	
8. Peers – Overall	.664**	.645**	.639**	.655**	.965**	.970**	.980**	1

Note: ** $p < 0.01$

Comparison of Peer Assessment and Expert Assessment

As a basis for the later simulation study, we compared the assessment carried out by the peers with the assessment carried out by the experts in a first analysis. We checked the assumption for statistical tests and tested on standard normal distribution for our data. A Kolmogorov-Smirnov test revealed that the data for expert assessors are normally distributed and the data for the peer assessors are not. Therefore, the spearman rank correlation coefficient was used to study the relationships between the peer assessors and the expert assessors. Table 1 shows the results.

The results show high significance between the peer and the expert assessors; meaning that both measure the same. Moreover, the correlation between the peer and the expert assessors is of large characterization (Cohen 1988). Hence, we can show that decisions of both peer and the expert assessors are related. Further, we used a meaningful statistical evaluation to compare both groups (peer and expert assessors) regarding their assessment. We used the non-parametric Mann-Whitney U-Test to compare both groups. Table 2 shows the results. For the three experts we use the abbreviation E1, E2, and E3. For the peers, we use the abbreviation P1, P2, P3, P4, and P5.

Table 2. Evaluation Results Regarding the Group Comparison for Independent Samples and the Effect Size

Variables	Expert (E1/E2/E3)	Peers (P1/P2/P3/P4/P5)	Effect Size d_{Cohen}	Mann–Whitney Z
Sample size	N = 136	N = 136		

Knowledge Expertise	Mean = 4.00
	S.D. = 1.15
		Mean = 5.69
		S.D. = .72
			d = 1.762	-11.281***

Solutions Quality	Mean = 4.06
	S.D. = 1.21
		Mean = 5.71
		S.D. = .78
			d = 1.621	-10.692***

Educational Objectives	Mean = 4.03
	S.D. = 1.16
		Mean = 5.61
		S.D. = .80
			d = 1.586	-10.532***

Overall	Mean = 4.03
	S.D. = 1.16
		Mean = 5.67
		S.D. = .76
			d = 1.672	-10.901***

Note: S.D. Standard Deviation. *** p < 0.001

The Mann-Whitney U-Test reveals that the difference of the mean values achieved in each variable and for all variables altogether (namely overall) is significant at p < .001. Thus, both groups are different in the mean values regarding the three variables knowledge expertise, solutions quality, and educational objectives as well as overall. Moreover, we gathered the effect size using Cohen's d which shows a great effect of the differences of the mean values in both groups (Cohen 1988). Thus, the effect size d underpins our observed effects regarding the difference of both groups. Considering the mean values of both groups, they reveal that the peers assess on higher values over all three variables and in overall compared to the experts. Therefore, it can be state that the assessments carried out by the peers are less strict compared to the assessments carried out by the three independent experts.

To sum up, we are able to show that both peer and expert assessors are of large correlation and the peers assess more positive than the experts.

Simulation results: Examining the Number of Peer Assessors

Within the second research question, we want to investigate how many students are required for a stable assessment on a peer's solution. Research investigation to examine a required number of users or responses to reach a stable assessment is of wide interest in various field of studies. For example, in online innovation communities participation usually fluctuates greatly which makes it necessary to control rating scale users for valuable contributions (Preece and Shneiderman 2009). The research study by Riedl et al. (2013) shows that when engaging consumers in co-creating products and services within online innovation communities about 20 user ratings per user idea are adequate to create a stable ranking of a user's idea. It is obvious that grounding the research study on a single respondent is more convenient (Van Bruggen et al. 1999); several research studies determined a multiple respondent-based approach leading to results with superior quality (Hill 1982; Wilson and Lilien 1992). Therefore, recommendations should rely on a multiple respondent-based approach for the research study (Van Bruggen et al. 1999). The aggregation of a smaller amount of obtainable ratings reduces the quality of the aggregated rating (Riedl et al. 2013) and as a consequence, outliers have a stronger weight (Rushton et al. 1983).

To answer the second research question we applied a bootstrap-based Monte Carlo simulation based on our data to investigate how many peer assessors are required to reach a stable assessment for another peer's solution. Monte Carlo simulations are a broad class of computational algorithms that count on repeated random sampling to receive numerical results (Kalos and Whitlock 2008; Rubinstein and Kroese 2011). There are several research studies using a Monte Carlo and bootstrap simulations, for example in the field of information systems (Goodhue et al. 2007; Riedl et al. 2013), natural sciences (El-adaway 2011; Huelsenbeck and Ronquist 2001), marketing research (Vriens et al. 1996), and financing (Jonsson and Lindbergh 2013). The bootstrapping method belongs to the resampling methods and allows the approximation of parameters based on N replications from the original data set (Efron and Tibshirani 1993).

We used a resampling-based simulation to approximate the peer assessment to the aggregate expert assessment. As aggregate assessment we used the mean values of each variable and for the variables overall. In the resampling-based simulation we randomly drew assessments with replacement from the data set of expert assessments and aggregated these assessments. We then calculated the root mean square error (RMSE) and the mean absolute percentage error (MAPE) for error measurement to investigate the deviation of the results of the peer assessment compared to the expert assessment. The results for RMSE and MAPE are similar. Therefore, we only report the data for RMSE at this point.

The RMSE is regularly employed in model evaluation studies (Chai and Draxler 2014) and presents the difference between predicted values and observed values. When the sample size is above 100, reconstructing the error distribution using RMSEs is even more reliable (Chai and Draxler 2014). In general, the RMSE is defined by the following equation (Eq):

$$Eq. 1: RMSE = \sqrt{1/n \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

The RMSE reveals to which extent the estimated value $\hat{y}_i$ deviates from the measured value y_i. In our case, the RMSE allows a relative comparison between the peer assessors and the expert assessors. The calculation for the RMSE is based on the items regarding our rating sheet in the assessment form. Each item could be assessed on a scale ranging from 1 to 7 (cf. section data collection). This leads to an error metric ranging from 0 to 6 for the RMSE, whereby 0 means complete agreement and 6 means complete deviation. For measuring the RMSE item-based, the value for the n is the result of our sample size from 136 students and the 18 items (hence, the n is 2448). For our research study, the RMSE is defined as follows:

$$Eq. 2: RMSE = \sqrt{1/n \sum_{i=1}^{n} (expert\ assessment - \ peer\ assessment)^2}$$

Figure 2 shows the Monte Carlo approximation of 1000 simulations that runs for each variables and the RMSE. The figure shows that all variables start at a high RMSE after using one out of five peers (RMSE of 1.88 for knowledge expertise; RMSE of 1.88 for solution quality; RMSE of 1.82 for educational objectives; RMSE of 1.86 over all three variables). The values drop sharply with an additional peer assessor until a peer assessor three out of five is achieved. Then, all curves level off after three peer assessors. The difference in the RMSE from four peer assessors compared to three and from five peer assessors compared to four is marginal.

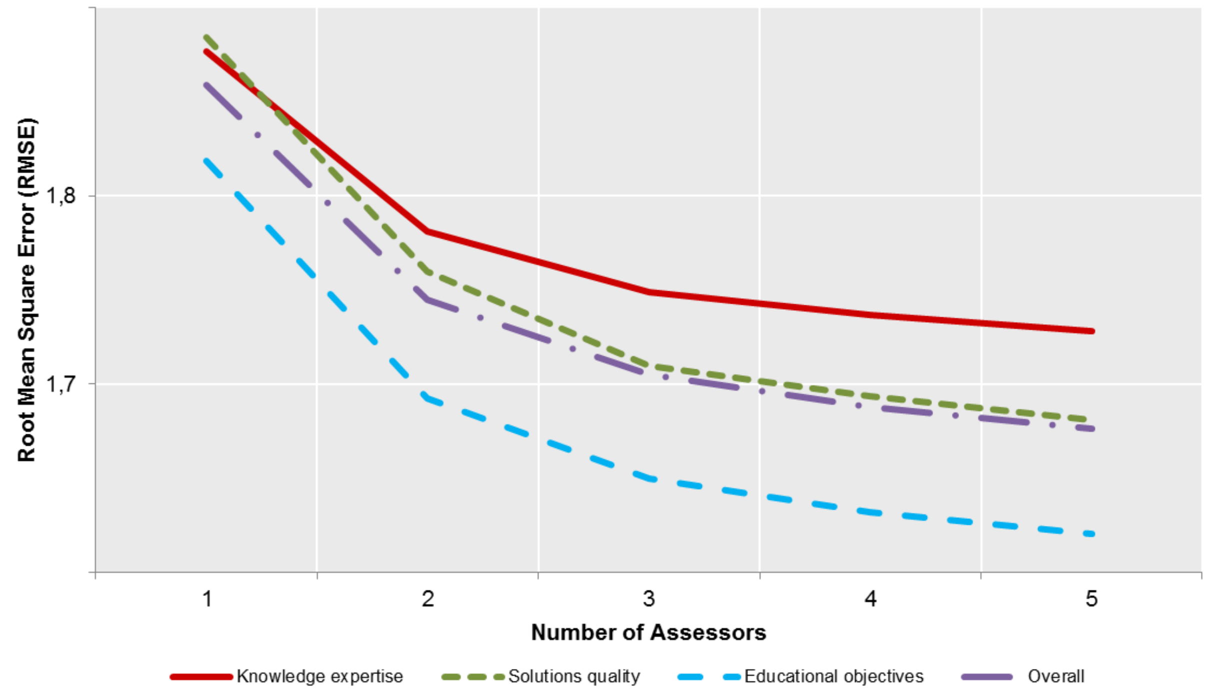

Figure 2. Accuracy of Peer Assessment by Dimension. Simulation Results.

The following table 3 shows in detail the values for the RMSE that are achieved when randomly one peer assessor out of five was drawn, two peer assessors out of five, and so on. Furthermore, the results for RMSE are reported for each variable and over all three variables. Additionally, we calculated the performance gain for each variable and overall. The performance gain shows the percentage deviation of the RMSE from a specific number of peer assessors to one peer assessor less.

As the results in table 3 shows in specific, the difference in the RMSE between one and two peer assessors as well as two and three peer assessors is larger compared to the difference between three and four peer assessors as well as four and five peer assessors. Additionally, the difference for RMSE and performance gain between two and three peer assessors is larger compared to the difference between three and five peer assessors. Furthermore, the curves in figure 2 show the main difference between the drop and the level off after the use of three peer assessors. Therefore, the results of the performance gain in table 3 prove the curves. Thus, it can be concluded that three of five peer assessors are required for a stable assessment on a peer's solution. The marginal difference to the feedback after four or five peers is not in relation to the expense and effort students have by providing feedback on four or five peers.

Table 3. Values of RMSE and Performance Gain

Number Peer Assessor	RMSE KE	PG for KE	RMSE SQ	PG for SQ	RMSE EO	PG for EO	RMSE Overall	PG for Overall
1	1.88		1.88		1.82		1.86	
2	1.78	5.62%	1.76	6.82%	1.69	7.69%	1.74	6.90%
3	1.75	1.71%	1.71	2.92%	1.65	2.42%	1.71	1.75%
4	1.74	.57%	1.69	1.18%	1.63	1.23%	1.69	1.18%
5	1.73	.58%	1.68	.60%	1.62	.62%	1.68	.60%

Note: KE Knowledge expertise, SQ Solutions quality, EO Educational objectives, PG Performance gain.

Discussion

Although the use of peer assessment is not new in higher education, no research studies exist that investigate how many peer assessors are required for a stable assessment on a peer's solution concerning the use of complex free-text assignments addressing the high cognitive levels of educational objectives. Therefore, the aim of this study was to examine how many peers as assessors are required to provide a stable assessment on a peer's solution. Moreover, we sought to investigate the comparability of the assessments carried out by peer assessors and the assessments carried out by expert assessors.

The results of the first research question show that the assessment by peers and experts are different from each other. Specifically, the students assess the solutions of their peers more positive compared to the assessment carried out by the experts. One reason could be that the students are more lenient in the judgement of their fellow students. The lenient judgement might be the consequence that the students were in their first or second year of study and had no experiences in assessing their peers. In this lecture it was the first time that we conducted the peer assessment. Moreover, we used a complex free-text assignment for the first time which addressed high cognitive levels of educational objectives. These were many innovations for the students. So, it might be possible that the students acting as assessors were uncertain in their role. On the one hand, it might be possible that the students were uncertain in providing feedback to their fellow students and they did not want to be strict, even if it is anonymously. On the other hand, it might be possible that the students perceived it as difficult to assess the performance on a peers' solution to a complex free-text assignment. Hence, these might be explanations for the lenient assessment by the peers among each other. This social phenomenon, not to provide a negative judgement to other persons as well as to tend to heap the judgement at the positive end of a scale (Tourangeau et al. 2000) belongs to the response bias in psychological literature (Podsakoff et al. 2003). The positive judgement of a respondent of the performance, attitudes, or perceptions of others is known as the positivity or leniency bias (Podsakoff et al. 2003). This tendency for a lenient judgment appears in performance evaluation of employees (Landy and Farr 1980), for example when managers intentional change appraisals in order to support employees, to improve training or to avoid tensions (Levy and Williams 2004; Prendergast 2002). Moreover, lenient evaluations are also known in evaluations of political figures (Lau et al. 1979), as well as in ratings of educational settings (Sears 1983), when students presume to receive a more positive grade by a lenient judgment of lectures and lecturers for example (Marsh and Roche 2000). Prior research studies show that the confidence in providing feedback and feedbacks' quality will improve with more practice (Brutus and Donia 2010; Gielen and De Wever 2015). Therefore, we expect to achieve results which show higher comparability between peer and expert assessors when the students are more confident and more experienced in assessing their peers.

Another reason for the deviation between peer and expert assessment might be that we conducted the peer assessment anonymously. Of course, we did this intentionally according to relevant peer assessment literature to avoid manipulations concerning social relationships and to ensure objective feedback modeling (Boud and Falchikov 2007). It is stated in literature that students underestimate the work of others they do not like and they overestimate the work of friends (Bostock 2004). However, being uncertain whose solution oneself is evaluating could be another reason for providing feedback less strict. The result of our first research question showing the deviation between peer and expert assessors raises of course the legitimate question whether a peer assessment is useful to apply as a didactical method in education. Answering this question leads to another question, namely what is an alternative? One alternative could be that students did not receive any feedback concerning their current learning progress during the teaching-learning process. Anyway, there are many research studies showing that it is highly relevant and valuable providing students with formatively individual feedback leading, among others, to more students' satisfaction (Rubin et al. 2010), higher self-regulation (Gielen et al. 2010), and higher learning outcome (Hsia et al. 2016). Therefore, not using peer assessment is not an option. We are convinced that receiving feedback from peers helps the students a lot, even if the feedback is more lenient and higher rated than a feedback from experts. As the literature shows, students receive a deeper understanding of the learning content, achieve metacognitive skills, and gain new ideas for further improvement of the own solution by reading the solutions of others and by providing feedback to others (Jaillet 2009; Sadler and Good 2006). Moreover, students are getting aware of their own strengths and weaknesses and are able to self-monitor the own learning progress (Tahir 2012). Hence, the peer assessment is a suitable didactical method to train several skills and competencies for the individual student and therefore, goes a step further than the merely performance assessment.

Especially in large-scale lectures, didactical methods are needed to antagonize the challenges these lectures are faced with, such as the lack of interaction and feedback. Regarding our results, by using peer assessment a possibility might be that the lecturer informs the students about scientific results that students provide a lenient assessment among each other. Thus, the students will know that the received feedback is more positive compared to the feedback the lecturer would provide. This might be helpful for the students to self-assess the received feedback from the peers. Another option might be that the students receive the qualitative feedback only and regarding the quantitative feedback the students will be informed whether they are better or worse than the average of all students. This is feasible with the LMS Moodle, but makes additional resources from the lecturer necessary.

Regarding the results of the second research question, by using a bootstrap-based simulation we were able to show that an average of three peer assessors for providing feedback on another peer's solution leads to a stable assessment. Hence, three peer assessors are required for the use of a peer assessment. Adding more peer assessors for providing feedback increases the accuracy only very slightly (cf. the results in table 3 and the curves in Figure 2figure 2). Moreover, the knowledge benefit of the additional received feedback of four or five peers is not in relation to the expense and effort. Furthermore, using three instead of four or five peer assessors facilitate the work for the peer assessors. By providing feedback to three other peers allows the peer assessor to work more conscientious and more precise compared to the effort that must be spent by providing feedback of the same quality to four or five peers. In turn, a more conscientious and more precise feedback increases the benefit for the peer assesse.

With the result how many assessors are required for a stable peer assessment the lecturer is supported regarding the specific design decision when using peer assessment in the lecture. The result that three peer assessors are required increases the quality for peer assessment overall. The lecturer has access to the solution of each peer and is therefore able to monitor and intervene in the teaching-learning process whenever necessary. Thus, misunderstandings and relevant ambiguities can be eliminated and content-specific questions discussed. This increases the lecture quality and emphasizes the learning process of technology-mediated lectures (Söllner et al. 2017).

Theoretical and Practical Implications

For the theoretical implication, the study contributes to the involvement as well as the emphasis of the role of peers in the teaching-learning process. The learners adopt the lecturer's role and assess their peers, which can enable reflection and metacognition. The results provide insights on how to integrate a learner-centered approach which includes the integration of interaction and feedback in large-scale lectures. Moreover, the study underlines pedagogical research regarding imparting and verification of knowledge on the higher cognitive levels of learning objectives in large-scale lectures at an undergraduate level. This enables a complex and comprehensive understanding of the course content beyond the factual knowledge.

An application of the results to other teaching-learning environments is also possible. Peer assessment is a valuable and enhancing learning approach not only for large-scale blended learning lectures, but also for e-learning lectures (e.g., MOOCs), or traditional teaching-learning environments, or other new contexts of interest.

The results are of scientific and practical relevance in terms of education, since they provide insights on how to integrate interaction and feedback into the teaching-learning process by means of peer assessment, and at the same time antagonize the challenges of large-scale lectures. The study reveals a way to engage and activate students in large-scale lectures in spite of limited resources. The results comply with the lecturer's limited resources and do not additionally increase their workload. Our study offers useful insights into helping learners pursue their own learning progress. It suggests that higher levels of formative assessment are desirable, as they strongly demand learners' awareness for self-regulation of their own learning progress.

Despite the difference between the peer and the expert assessment, the paper contributes in trusting peer assessment as didactical method in teaching. The results show that the peers provide a higher assessment regarding all three variables used in the rating sheet. But, the deviation between the peers and the experts is steady. Moreover, the results of this research study contribute to the peer assessment research. In specific, the results offer insights that three peers are required to provide a stable assessment on a peers' solution. This answers an economic question regarding the number of peers in a peer assessment, which is highly relevant for other researchers using peer assessment. Moreover, this result contributes to the design of the peer assessment. Furthermore, the result regarding the sufficient number of three peer assessors can be extended to further research (for example assessing a prototype in the product development or assessing (business) ideas in open innovation scenarios).

Conclusions, Limitations and Future Research

Any findings or implications of this study need to be considered in light of its limitations. When using peer assessment, the students need particular skills and abilities (Gielen and De Wever 2015). We provide the students with specific assessment rules, which show how to provide and receive feedback. However, we used the peer assessment for the first time in our lecture and the undergraduate students had no previously experiences in assessing their peers. Therefore, it might be possible that the students acting as assessors were uncertain in providing feedback and assessing a peers' solution to a specific assignment. We need to take into consideration that this could limit our results regarding the first research question. Consequently, future research could conduct a comparable experiment, e.g., with graduate students, and see whether the results differ.

Another limitation addresses the complex free-text assignment we used for the peer assessment. We need to be aware that other results might be achieved with another assignment type. By using less complex assignments, such as single/multiple choice or true/false statements, it might be possible to achieve an agreement between the peer and the expert assessment. Assignments with a clear solution apparent by checking a box make assessment easier than assessing a written text with argumentation and discussion.

One change we will implement in the peer assessment is the frequency of use. Currently, we used the peer assessment once in our lecture. The idea is to repeat the peer assessment process several times during a lecture. The students will then train their ability in providing feedback and critical thinking. We then assume to achieve more similar results for the peer assessors which are more consistent and reliable to the expert assessors'. This could be investigated in a future research study. Another change we want to undertake is that each student has to rate the feedback they receive from their peers. The aim is to investigate how the students' skills in providing feedback will change by using peer assessment several times during the teaching-learning process.

The findings of our research are based on the feedback the peers provide among each other in quantitative format. Additionally, the peer assessment captured the feedback in qualitative formats. Future research could investigate how qualitative feedback needs to be designed to be of high quality and therefore, of high reliability and validity. Moreover, it could be interesting to examine how many students are required for a stable assessment considering the qualitative feedback. To answer these research questions, the peer assessment should be conducted in a setting with a smaller number of participants.

Another option for future research investigation refers to the use of the assessment form. In our study we used the quantitative rating sheet regarding several criteria for data analysis. Instead of using criteria it could be possible to conduct the assessment with credit points respectively under the assumption of providing credit points. This would be pursuant to the procedure in the final exam where students receive credit points on their performance. Moreover, future research should conduct the peer assessment with students in the advanced part of their studies; namely in masters seminars for example. At this point of study, students should be more experienced in solving complex free-text assignments on their own. Therefore, it would be interesting to investigate how peer and expert assessors interrelate.

Acknowledgements

The research presented in this paper was partially funded by the German Federal Ministry of Education and Research in the course of the project StaySmart (http://projekt-staysmart.de), Grant No. 01FK14008. The responsibility for the content of this publication lies with the authors. Our thanks go to our student assistant Andreas Gibhardt for his thorough and valuable support during the development process of the paper.

References

Amabile, T. 1996. Creativity in Context. Westview press.

Anderson, L., Krathwohl, D., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P., and Wittrock, J. 2001. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. New York: Addison Wesley Longmann.

Ball, D. M., and Levy, Y. 2008. "Emerging Educational Technology: Assessing the Factors That Influence Instructors' Acceptance in Information Systems and Other Classrooms," Journal of Information Systems Education (19:4), pp. 431-444.

Bauer, C., Figl, K., Derntl, M., Beran, P. P., and Kabicher, S. 2009. "Der Einsatz Von Online-Peer-Reviews Als Kollaborative Lernform," Wirtschaftsinformatik (2)), pp. 421-430.

Biemer, P., and Lyberg, L. E. 2003. Introduction to Survey Quality, (1. ed.). Hoboken, NJ, USA: John Wiley & Sons.

Bligh, D. 2000. What's the Use of Lectures? San Francisco: Jossey-Bass Publisher.

Bloom, B. S., and Krathwohl, D. R. 1956. Taxonomy of Educational Objectives. The Classification of Educational Goals, by a Committee of College and University Examiners. Handbook I, Cognitive Domain. New York: Green Longmans

Bostock, S. J. 2004. "Motivation and Electronic Assessment," Effective Learning and Teaching in Computing), pp. 86-99.

Boud, D., and Falchikov, N. 2007. Rethinking Assessment in Higher Education: Learning for the Longer Term. Abingdon, Oxon: Routledge.

Boydell, D. 1994. "The Use of Peer Group Review in the Assessment of Project Work in Higher Education," Mentoring & Tutoring: Partnership in Learning (2:2), pp. 45-52.

Braun, E., Gusy, B., Leidner, B., and Hannover, B. 2008. "Das Berliner Evaluationsinstrument Für Selbsteingeschätzte, Studentische Kompetenzen (Bevakomp)," Diagnostica (54:1), pp. 30-42.

Brindley, C., and Scoffield, S. 1998. "Peer Assessment in Undergraduate Programmes," Teaching in Higher Education (3:1), pp. 79-90.

Brutus, S., and Donia, M. B. 2010. "Improving the Effectiveness of Students in Groups with a Centralized Peer Evaluation System," Academy of Management Learning & Education (9:4), pp. 652-662.

Chai, T., and Draxler, R. R. 2014. "Root Mean Square Error (Rmse) or Mean Absolute Error (Mae)?– Arguments against Avoiding Rmse in the Literature," Geoscientific Model Development (7:3), pp. 1247-1250.

Chang, C.-C., Tseng, K.-H., and Lou, S.-J. 2012. "A Comparative Analysis of the Consistency and Difference among Teacher-Assessment, Student Self-Assessment and Peer-Assessment in a Web-Based Portfolio Assessment Environment for High School Students," Computers & Education (58:1), pp. 303-320.

Chen, C.-h. 2010. "The Implementation and Evaluation of a Mobile Self-and Peer-Assessment System," Computers & Education (55:1), pp. 229-236.

Cheng, W., and Warren, M. 1999. "Peer and Teacher Assessment of the Oral and Written Tasks of a Group Project," Assessment & Evaluation in Higher Education (24:3), pp. 301-314.

Cohen, J. 1988. Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum.

Conway, J. M., and Huffcutt, A. I. 1997. "Psychometric Properties of Multisource Performance Ratings: A Meta-Analysis of Subordinate, Supervisor, Peer, and Self-Ratings," Human Performance (10:4), pp. 331-360.

Darling-Hammond, L., Ancess, J., and Falk, B. 1995. Authentic Assessment in Action: Studies of Schools and Students at Work. Teachers College Press.

Dochy, F., Segers, M., and Sluijsmans, D. 1999. "The Use of Self-, Peer and Co-Assessment in Higher Education: A Review," Studies in Higher Education (24:3), pp. 331-350.

Efron, B., and Tibshirani, R. J. 1993. An Introduction to the Bootstrap. New York Chapman & Hall.

El-adaway, I. H. 2011. "Insurance Pricing for Windstorm-Susceptible Developments: Bootstrapping Approach," Journal of Management in Engineering (28:2), pp. 96-103.

Eveland, W. P., Jr., and Sharon, D. 2000. "Examining Information Processing on the World Wide Web Using Think Aloud Protocols " Media Psychology (2), pp. 219-244.

Falchikov, N., and Goldfinch, J. 2000. "Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks," Review of Educational Research (70:3), pp. 287-322.

Fortes, P. C., and Tchantchane, A. 2010. "Dealing with Large Classes: A Real Challenge," Procedia - Social and Behavioral Sciences (8:0), pp. 272-280.

Gardner, J., Scogin, F., Vipperman, R., and Varela, J. G. 1998. "The Predictive Validity of Peer Assessment in Law Enforcement: A 6 Year Follow-Up," Behavioral Sciences & the Law (16:4), pp. 473-478.

Gielen, M., and De Wever, B. 2015. "Scripting the Role of Assessor and Assessee in Peer Assessment in a Wiki Environment: Impact on Peer Feedback Quality and Product Improvement," Computers & Education (88), pp. 370-386.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., and Struyven, K. 2010. "Improving the Effectiveness of Peer Feedback for Learning," Learning and Instruction (20:4), pp. 304-315.

Goodhue, D., Lewis, W., and Thompson, R. 2007. "Research Note-Statistical Power in Analyzing Interaction Effects: Questioning the Advantage of Pls with Product Indicators," Information Systems Research (18:2), pp. 211-227.

Graesser, A. C., Pearson, N. K., and Magliano, J. P. 1995. "Collaborative Dialogue Patterns in Naturalistic One-to-One Tutoring," Applied Cognitive Psychology (9), pp. 495-522.

Greguras, G. J., and Robie, C. 1998. "A New Look at within-Source Interrater Reliability of 360-Degree Feedback Ratings," Journal of Applied Psychology (83:6), p. 960.

Hafner, J., and Hafner, P. 2003. "Quantitative Analysis of the Rubric as an Assessment Tool: An Empirical Study of Student Peer-Group Rating," Int. J. Sci. Educ. (25:12), pp. 1509-1528.

Hagstrom, F. 2006. "Formative Learning and Assessment," Communication Disorders Quarterly (28:1), pp. 24-36.

Harter, S. P. 1986. Online Information Retrieval: Concepts, Principles and Techniques. Orlando, FL: Academic Press.

Hill, G. W. 1982. "Group Versus Individual Performance: Are N+ 1 Heads Better Than One?," Psychological Bulletin (91:3), p. 517.

Hovardas, T., Tsivitanidou, O. E., and Zacharia, Z. C. 2014. "Peer Versus Expert Feedback: An Investigation of the Quality of Peer Feedback among Secondary School Students," Computers & Education (71), pp. 133-152.

Hsia, L.-H., Huang, I., and Hwang, G.-J. 2016. "Effects of Different Online Peer-Feedback Approaches on Students' Performance Skills, Motivation and Self-Efficacy in a Dance Course," Computers & Education).

Huelsenbeck, J. P., and Ronquist, F. 2001. "Mrbayes: Bayesian Inference of Phylogenetic Trees," Bioinformatics (17:8), pp. 754-755.

Hughes, I. E., and Large, B. J. 1993. "Staff and Peer-Group Assessment of Oral Communication Skills," Studies in Higher Education (18:3), pp. 379-385.

Jaillet, A. 2009. "Can Online Peer Assessment Be Trusted?," Educational Technology & Society (12:4), pp. 257-268.

Jonsson, S., and Lindbergh, J. 2013. "The Development of Social Capital and Financing of Entrepreneurial Firms: From Financial Bootstrapping to Bank Funding," Entrepreneurship Theory and Practice (37:4), pp. 661-686.

Kalos, M. H., and Whitlock, P. A. 2008. Monte Carlo Methods. John Wiley & Sons.

Knight, J. K., and Wood, W. B. 2005. "Teaching More by Lecturing Less," Cell biology education (4:4), pp. 298-310.

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., and Klemmer, S. R. 2013. "Peer and Self Assessment in Massive Online Classes," ACM Transactions on Computer-Human Interaction (20:6), p. 33.

Landy, F. J., and Farr, J. L. 1980. "Performance Rating," Psychological Bulletin (87:1), p. 72.

Lau, R. R., Sears, D. O., and Centers, R. 1979. "The "Positivity Bias" Inevaluations of Public Figures: Evidence against Instrument Artifacts," Public Opinion Quarterly (43:3), pp. 347-358.

Lehmann, K., and Leimeister, J.-M. 2015. "Theory-Driven Design of an It-Based Peer Assessment to Assess High Cognitive Levels of Educational Objectives in Large-Scale Learning Services," 23rd European Conference on Information Systems (ECIS 2015), Münster, Germany.

Lehmann, K., and Söllner, M. 2014. "Theory-Driven Design of a Mobile-Learning Application to Support Different Interaction Types in Large-Scale Lectures," in: European Conference on Information Systems (ECIS). Tel Aviv, Israel.

Lehmann, K., Söllner, M., and Leimeister, J. M. 2016. "Design and Evaluation of an It-Based Peer Assessment to Increase Learner Performance in Large-Scale Lectures," International Conference on Information Systems (ICIS), Dublin, Ireland.

Leijen, Ä., Lam, I., Wildschut, L., Simons, P. R.-J., and Admiraal, W. 2009. "Streaming Video to Enhance Students' Reflection in Dance Education," Computers & Education (52:1), pp. 169-176.

Levy, P. E., and Williams, J. R. 2004. "The Social Context of Performance Appraisal: A Review and Framework for the Future," Journal of management (30:6), pp. 881-905.

Lin, S. S. J., Liu, E. Z.-F., and Yuan, S.-M. 2001. "Web-Based Peer Assessment: Feedback for Students with Various Thinking-Styles," Journal of Computer Assisted Learning (17:4), pp. 420-432.

Marsh, H. W., and Roche, L. A. 2000. "Effects of Grading Leniency and Low Workload on Students' Evaluations of Teaching: Popular Myth, Bias, Validity, or Innocent Bystanders?," Journal of Educational Psychology (92:1), p. 202.

Mathews, B. P. 1994. "Assessing Individual Contributions: Experience of Peer Evaluation in Major Group Projects," British Journal of Educational Technology (25:1), pp. 19-28.

Mowl, G., and Pain, R. 1995. "Using Self and Peer Assessment to Improve Students' Essay Writing: A Case Study from Geography," Programmed Learning (32:4), pp. 324-335.

O'Neill, T. A., Goffin, R. D., and Gellatly, I. R. 2012. "The Knowledge, Skill, and Ability Requirements for Teamwork: Revisiting the Teamwork-Ksa Test's Validity," International Journal of Selection and Assessment (20:1), pp. 36-52.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., and Podsakoff, N. P. 2003. "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," Journal of applied psychology (88:5), p. 879.

Preece, J., and Shneiderman, B. 2009. "The Reader-to-Leader Framework: Motivating Technology-Mediated Social Participation," AIS Transactions on Human-Computer Interaction (1:1), pp. 13-32.

Prendergast, C. 2002. "Uncertainty and Incentives," Journal of Labor Economics (20:S2), pp. S115-S137.

Riedl, C., Blohm, I., Leimeister, J. M., and Krcmar, H. 2013. "The Effect of Rating Scales on Decision Quality and User Attitudes in Online Innovation Communities," International Journal of Electronic Commerce (17:3), pp. 7-36.

Rubin, B., Fernandes, R., Avgerinou, M. D., and Moore, J. 2010. "The Effect of Learning Management Systems on Student and Faculty Outcomes," The Internet and Higher Education (13:1–2), pp. 82-83.

Rubinstein, R. Y., and Kroese, D. P. 2011. Simulation and the Monte Carlo Method. John Wiley & Sons.

Rushton, J. P., Brainerd, C. J., and Pressley, M. 1983. "Behavioral Development and Construct Validity: The Principle of Aggregation," Psychological Bulletin (94:1), pp. 18-38.

Sadler, P. M., and Good, E. 2006. "The Impact of Self-and Peer-Grading on Student Learning," Educational Assessment (11:1), pp. 1-31.

Santhanam, R., Sasidharan, S., and Webster, J. 2008. "Using Self-Regulatory Learning to Enhance E-Learning-Based Information Technology Training," Information Systems Research (19:1), pp. 26-47.

Schumacher, J. E., Scogin, F., Howland, K., and McGee, J. 1992. "The Relation of Peer Assessment to Future Law Enforcement Performance," Criminal justice and behavior (19:3), pp. 286-293.

Sears, D. O. 1983. "The Person-Positivity Bias," Journal of Personality and Social Psychology (44:2), p. 233.

Shrout, P. E., and Fleiss, J. L. 1979. "Intraclass Correlations: Uses in Assessing Rater Reliability," Psychological bulletin (86:2), p. 420.

Söllner, M., Bitzer, P., Janson, A., and Leimeister, J. M. 2017. "Process Is King: Evaluating the Performance of Technology-Mediated Learning in Vocational Software Training.," Journal of Information Technology (JIT)), pp. 1-21.

Söllner, M., Hoffmann, A., and Leimeister, J. M. 2016. "Why Different Trust Relationships Matter for Information Systems Users," European Journal of Information Systems (EJIS) (3:25), pp. 274-287.

Strijbos, J.-W., Ochoa, T. A., Sluijsmans, D. M., Segers, M. S., and Tillema, H. H. 2009. "Fostering Interactivity through Formative Peer Assessment in (Web-Based) Collaborative Learning Environments," Cognitive and emotional processes in web-based education: Integrating human factors and personalization), pp. 375-395.

Tahir, I. H. 2012. "A Study on Peer Evaluation and Its Influence on College Esl Students," Procedia - Social and Behavioral Sciences (68:0), pp. 192-201.

Topping, K. J. 1998. "Peer Assessment between Students in Colleges and Universities," Review of educational research (68:3), pp. 249-276.

Topping, K. J., Smith, E. F., Swanson, I., and Elliot, A. 2000. "Formative Peer Assessment of Academic Writing between Postgraduate Students," Assessment & evaluation in higher education (25:2), pp. 149-169.

Tourangeau, R., Rips, L. J., and Rasinski, K. 2000. The Psychology of Survey Response. Cambridge University Press.

Tseng, S.-C., and Tsai, C.-C. 2007. "Online Peer Assessment and the Role of the Peer Feedback: A Study of High School Computer Course," Computers & Education (49:4), pp. 1161-1174.

Tsivitanidou, O. E., Zacharia, Z. C., and Hovardas, T. 2011. "Investigating Secondary School Students' Unmediated Peer Assessment Skills," Learning and Instruction (21:4), pp. 506-519.

Umar, I. N., and Hui, T. H. 2012. "Learning Style, Metaphor and Pair Programming: Do They Influence Performance?," Procedia - Social and Behavioral Sciences (46:0), pp. 5603-5609.

Van Bruggen, G. H., Lilien, G. L., and Kacker, M. 1999. "Informants in Organizational Marketing Research: How Many, Who, and How to Aggregate Opinions?,").

Van Lehn, K. A., Chi, M. T. H., Baggett, W., and Murray, R. C. 1995. Progress Report: Towards a Theory of Learning During Tutoring. Pittsburgh, PA: Learning Research and Development Center, University of Pittsburgh.

Vriens, M., Wedel, M., and Wilms, T. 1996. "Metric Conjoint Segmentation Methods: A Monte Carlo Comparison," Journal of Marketing Research), pp. 73-85.

Wilson, E. J., and Lilien, G. L. 1992. "Using Single Informants to Study Group Choice: An Examination of Research Practice in Organizational Buying," Marketing Letters (3:3), pp. 297-305.

Appendix

Appendix A

Table A.1. Control Variables to Compare Treatment and Control Group (Lehmann et al. 2016)

Control Variables		Treatment Group			Control Group			p-value
	Sample Size	N = 129			N = 45			
Gender								
	male	44.96 % (n = 58)			53.33 % (n = 24)			0.333
	female	55.04 % (n = 71)			46.67 % (n = 21)			
Age		Mean	SD		Mean	SD		0.164
		23.25 years	2.61		23.78 years	2.43		
Course of Study								
	Business Studies	96.12 % (n = 124)			93.33 % (n = 42)			0.442
	Others	3.88 % (n = 5)			6.67 % (n = 3)			
	Number of Items (Source)	Mean	SD	Cronbach's Alpha	Mean	SD	Cronbach's Alpha	
Self-Efficacy for Self-Regulation	11 (Santhanam et al. 2008)	5.48	0.81	.857	5.35	0.88	.868	0.517
Technology Experience	3 (Ball and Levy 2008)	5.43	1.32	.884	5.58	1.33	.855	0.447

Significance with * p= < .05 / ** = p < .01 / *** = p < .001 SD = Standard Deviation

Table A.2. Comparing Treatment and Control Group Regarding Learning Performance (Lehmann et al. 2016)

Variables	Treatment Group	Control Group	t(df) = t-value
Sample Size	N = 129	N = 45	

Part 1:
Single-choice assignments (max. 10 points)
	Mean = 5.47
	SD = 1.66
		Mean = 5.36
		SD = 2.02
			t(172) = .384

Part 2:
Non peer-assessment related free-text assignments *(A2, A5)* (max. 13.5 points)
	Mean = 4.42
	SD = 3.26
		Mean = 3.98
		SD = 3.46
			t(172) = .779

Peer-assessment related free-text assignments *(A1,A3,A4)* (max. 20.5 points)	Mean = 7.35
	SD = 3.85
		Mean = 6.15
		SD = 3.49
			t(172) = 1.844*

Part 3:
BPMN assignments (max. 46 points)
	Mean = 23.39
	SD = 5.74
		Mean = 22.42
		SD = 6.46
			t(172) = .951

Significance with * p= < .05 SD = Standard Deviation