A comparative analysis of the consistency and difference among online self-, peer-, external- and instructor-assessments: the competitive effect

Abstract

In the last few years, self- and peer-assessment have been increasingly employed not only as an evaluation method, but also as a learning procedure. The consistency and difference between self- and peer-assessments as compared to instructor-assessments have been previously studied, and a friendship bias was discovered. In this study, we introduce external-assessment (products are assessed by students from a different university that are enrolled in a similar course), and compare self-, peer-, external- and instructor-assessments. The experience was conducted at two different universities separated by a significant distance, during two consecutive years, including a total of 97 students. At both universities, students developed websites and online tools were employed to organise the different types of assessments. The obtained results indicate that there is a high-level of consistency across the different kinds of assessments. Moreover, a competitive effect was discovered: students tended to award higher grades to students from their same university while they were harsher with the products from a distant university. From the learning perspective, and according to the students' final grade, the assessment experience correlated with learning gains.

Key words:
Self-assessment; peer-assessment; external-assessment; competitive effect; online-tools

1. Introduction

Interest in self- and peer-assessment has grown in the last few years (Li et al., 2015; Chang, Tseng, & Lou, 2012; Hovardas, Tsivitanidou, & Zacharia, 2014; Pereira, Echeazarra, Sanz-Santamaría, & Gutiérrez, 2014). Such methods are widely applied to encourage collaboration among students at the same university, and aim to strengthen students' role in the learning-teaching process (Llado et al., 2014). Massive Open Online Courses (MOOCs) also introduce these methods, both as learning and working procedures (Kulkarni et al., 2013). There are pedagogical reasons to incorporate these methods: they increase motivation, help in the development of evaluation competences, and foment learning based on the observation of peers' works (Llado et al., 2014; Lai & Hwang, 2015; Chen, Wei, Wua, & Uden, 2009). Moreover, there are practical reasons as well. These types of assessment can be used to provide sufficient high-quality feedback within a reasonable time when working with a considerable number of students (Kulkarni et al., 2013; Luo, Robinson, & Park, 2014). They can also contribute to the evaluation of individual work within a team (intra-group), professionalism in team-work, and the ability to work in a group (Falchikov, 2003, Kennedy, 2005; Willmot,

Pond, Loddington, & Palermo, 2008). Indeed, these methods have drawbacks as well: the cost of organising and supervising the peer-assessment process or students' lack of trust in peer-assessment (Llado et al., 2014; Hovardas et al., 2014; McGarr & Clifford, 2013).

Instructors' primary concern regarding self- and peer-assessment is the degree of agreement between their marks and those awarded by their students (Falchikov & Goldfinch, 2000; Toppings, 2003). This factor may restrict their use and, thus, deprive many students of its learning benefits. For that reason, several studies aim to analyse the quality of these kinds of assessments. Such studies address the validity of these valuations, i.e. consistency (expressed as a correlation), and differences among the assessors: self-, peer-, and instructor-assessment (Chang et al., 2012; Panadero, Romero, & Strijbos, 2013). In those studies, instructor ratings are assumed to be the gold-standard (Li et al., 2015). The main points of interest in these studies has been synthesised in several literature-review articles. For example, Toppings (1998) conducted a review of the literature on peer-assessment among students in higher education. The results in this review indicate that peer-assessment is of adequate validity in a wide variety of fields. This review was later extended by Toppings (2003) to include self-assessment. This study concluded that the validity of self-assessment tends to be a little lower and more variable. In 2000, a meta-analysis comparing peer and teachers marks in higher education was published (Falchikov & Goldfinch, 2000). Since then, computer-assisted peer assessment began to grow exponentially. Features such as on-line assignment submission, storage, communication and review management have been introduced (Li, Liu, & Steckelberg, 2010; Tseng & Tsai, 2007). A more recent meta-analysis (Li et al., 2015) focused on synthesizing findings from studies on peer assessment and included the work of Falchikov and Goldfinch (2000). Both reviews reached similar conclusions: peer-ratings generally show a high level of validity.

Instructors' fears regarding peer- and self-assessment are most likely due to the lack of certainty in terms of students' objectivity during the evaluation process. In this regard, some authors (Topping, 2010) have suggested that further experimental and quasi-experimental studies are necessary to contrast the effect of different variables, such as face-to-face versus distant peer-assessment. In addition, investigations into friendship (or enmity) effects and their potential for bias should also be conducted (Falchikov & Goldfinch, 2000). In this regard, two different models have been examined in the literature: grades given by face-to-face peers (Panadero et al., 2013), and grades assigned by distant peers, as in MOOC courses where peers do not know each other (Kulkarni et al., 2013). Nowadays, both models use computer-assisted methods to help in the assessment process, but the interactions among students, both before and after valuation, are quite different. In the face-to-face model, a bias associated with the friendship relationship among evaluators has been documented (Panadero et al., 2013), and it has also been observed that students dislike the fact that their assessors are also competitors (Lin, Liu, & Yuan, 2001). In the distant model, bias may be related to very

different backgrounds, knowledge and skills (Luo et al., 2014), student engagement (Kizilcec, Piech, & Schneider, 2013), or nationality (“patriotism” bias) (Kulkarni et al., 2013).

An alternative method consists of studying peer-assessment conducted among face-to-face and distant students of two different universities using computer-assisted methods. With such an alternative, students must be enrolled in similar courses, so that they have acquired equivalent competences to be assessed in the assignments. The valuation awarded by a student from a different university is called *external-peer-assessment* or just *external-assessment* (see Figure 1). In this model, it is clear whether students have previously met. The students from the same university know each other and may develop friendship or competitiveness; whereas, students from distant universities are unrelated, and friendship among them is unfeasible, but competitiveness is more likely. It is already a known fact that human beings tend to view their own (peer) group in a more favourable light (Legault & Amiot, 2014), and even when differences between groups are minimal and trivial, people tend to favour in-groups over out-groups (Pronin, 2006). This factor may play a role when a student or instructor is assessing the work of a student from his or her own university, and comparing it to the work of a student of a different university.

Offering a different perspective, the authors of (Boubouka & Papanikalaou, 2013) suggest that incorporating peer assessment into learning methods such as Project Based Learning (PBL) represents an interesting line of research. PBL is a learning method based on developing projects wherein students plan, implement, and evaluate projects that have real-world applications beyond the classroom. There are many benefits of PBL which are extensively documented in the literature: for instance, the possibility of connecting learning with reality, increasing motivation, and promoting problem-solving, among others (Tynjälä, Pirhonen, Vartiainen, & Helle, 2009; Domínguez & Jaime, 2010). Assessment in this context, rather than being conducted solely at the end of the course to measure results (summative assessment), should be conducted throughout the learning process (formative assessment) (Strijbos & Sluijsmans, 2010). This process entails that the students who conduct the assessment, while reviewing the work of their peers, also have the opportunity to reflect on their own work, taking note of their errors and weaknesses. In this way, staged project work (Søndergaard & Mulder, 2012) lends itself particularly well to integrated peer assessment. It allows feedback to be produced and digested for a project that is still in progress. The present study aims to promote learning by peer-observation as opposed to peer-feedback. According to (Chen et al., 2009), the former has a more significant influence on learning. Therefore, it is likely that such a positive impact also occurs when including external-assessment. Moreover, the differences in the given assessments, depending both on the quality of the assessed products and on the competency of the assessing student, should also be taken into consideration. The Dunning-Kruger effect (Kruger & Dunning, 1999) could arise in this context: wherein unskilled individuals tend to overestimate their abilities. And furthermore, the varying quality of the products can be considered in order to further

contextualise the agreement between the different marks awarded by the various points of view (Sadler & Good, 2006).

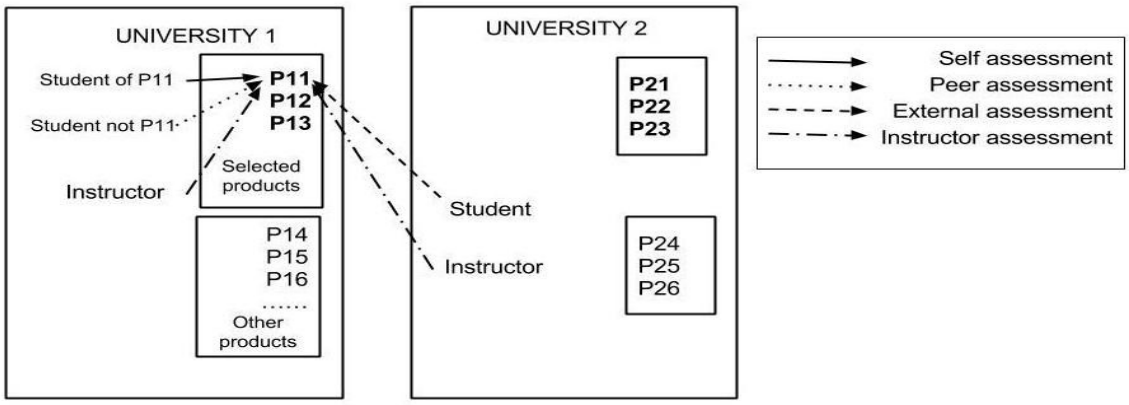

Fig. 1 Three kinds of assessments (self, peer, and external) depending on the social-implication of the student with the development team, and the instructor assessment.

The principal goal of this study consists in examining the consistency and the differences among self-, peer-, external- and instructor-assessment. To the best of our knowledge, the inclusion of external-assessment in this study represents an innovation in the literature. The reason to conduct such a study is to analyse whether the students valuate similarly their classmates' work and the external students' work. There are two additional, and secondary, goals: identify differences in the given assessments depending on both the quality of the assessed products and on the competency of the assessing student, and analyse the impact of this activity on the learning process.

We propose the following hypotheses:

1. There will be a high-level consistency across the different kinds of assessments.
2. Students will award better valuations to the products conducted at their own university as opposed to the products from the other university. Moreover, students will tend to choose the work created at their own university as the best.
3. Self-assessment will be less demanding than peer-assessment. And furthermore, external-assessment will be the strictest.
4. On average, instructor-assessments will be similar to peer-assessment, but there will be variations depending on (i) the quality of the assessed products, and (ii) the competency of the assessing student.
5. The activity will have a positive impact on student learning (measured by the final grade obtained in the course).

2. Material and methods

2.1. Research design

In order to test the aforementioned hypotheses, a study conducted during the 2013/14 and 2014/15 academic years was included. The context of this study is two introductory courses on computing project management in the computing engineering degree program at University 1 (U1) and University 2 (U2). Both courses share the same general goal, most of the competences to be developed, the number of credits, the level (third year), and the semester (second semester). Using a teaching-method based on PBL, students from each university worked in teams to create available web products that had to contain a video and verify some requirements: the requirements are the same for the products at both universities. Each team had to develop three different products (P1, P2 and P3) that were submitted sequentially during the course. Namely, products P1 from all the teams form a family and were submitted altogether (and likewise, for products P2 and P3). The products typically consisted of creating a web project using the computing project management techniques explained during the course. The website had to contain aspects from the students' university or city (for instance, an image of the faculty building) and also include the students' names. Therefore, the website clearly established the origin of its creators. After each family of products was submitted, the instructors picked three products from their respective university. Those selected products were assessed individually by all the students and instructors from both universities. The products were selected based on their qualities and as a sample of good and bad product. This selection facilitates comparison and simplifies the assessment of grades. Students do not know whether their products will be selected. The assessment was carried out after the deadline, but not long afterwards. Moreover, all the products from each family were assessed during the same session. Therefore, products were assessed and compared at the same time. In addition, students were asked to identify the best product. The assessment given by students did not influence the final grade of the valuated team or the assessing student (provided that the assessment was performed properly).

In the work presented in this paper, web systems have played an instrumental role. First of all, the products created by the students were usually websites including multimedia products (namely, videos) that are available online which facilitates their valuation. The availability of the products means that students are fully responsible for the product created – since the product may be examined by anyone, students are aware of their responsibility for the developed product. Secondly, the valuation was performed using web forms. Google-forms were utilized for the assessments in this study. Namely, instructors created a form for each set of valuated products. For each product, the form includes a page containing the link to the web resource and the assessment rubric – these forms are easy to develop, customize and publish. Several authors have also built web-based self- and peer- assessment support systems, see for instance (Li et al., 2010; Cho, Schunn, & Wilson, 2006). In our case, it was not necessary to develop a new

system, since the available technology could be employed. Nowadays, tools for communication and management of information systems offer interuniversity-teaching; moreover, most of them are free of cost (Jaime, Domínguez, Sánchez, & Blanco, 2013). It is estimated that each student or instructor devoted approximately 10 minutes to assessing each product. Finally, Google forms facilitated the dissemination and transparency of the results. In some cases, the synthesis of the collected assessments/evaluations was made available to students almost automatically after the valuation process – Google-forms collect the assessments in a database and allow the user to show the grouped (and anonymous) results for each valuated product. The aim was to show the students how their work is perceived by others.

In order to point out the peculiarities of the method applied herein, the inventory of peer assessment diversity suggested by (Gielen, Dochy, & Onghena, 2011) that completes the typology initially proposed by (Topping, 1998) was implemented. These characteristics are detailed in Table A.1 of the Appendix.

The chosen products were assessed using a simple rubric that contains three questions related to specific aspects of the work (satisfaction of requirements, video quality and web quality) and an overall valuation. For each question, a 5-point Likert scale ranging from 5 (very good) to 1 (very bad) was employed. A question asking for the best product in the assessed family was also included. Finally, students had to introduce an identification code; hence, assessments were not anonymous for the instructor (it is our belief that students should be able to justify their valuations).

A statistical analysis was also included in the study. In all the analyses, the conditions required for the parametric tests were verified. When these conditions were not met, it has been explicitly noted in this article, and then the corresponding non-parametric tests were used. In particular, ANOVA was employed with repeated measures to test whether there were differences among the three (self, peer and external) assessments methods. And secondly, each pair of methods was compared by means of a paired t-test using Bonferroni correction. Student’s t test was employed to check whether two sets of data were significantly different from each other. When parametric conditions were not verified, the corresponding non-parametric test (i.e. Mann-Whitney U test) were taken into account. The Pearson correlation coefficient was utilised to test the correlation between two variables. And finally, a chi-square test was employed to study the distribution independence for categorical data.

2.2. Sample

A total of 97 students participated in the study throughout two academic years, between 2013 and 2014, 82.5% of this group was male (48 from U1, 83.3% male; and 49 from U2, 82.6% male). The group of instructors consisted of 6 people, 83.3% male (2 from U1 and 4 from U2). The number of selected products was 36 (9 products per year and per university). In order to compare the effects on the learning-experience, data from the

previous academic year (2012) were included: the number of students during that academic year was 25, with 76% male (18 from U1 and 7 from U2). During the 2012 academic year, a similar teaching methodology was employed, but students' assessments were not included.

3. Results and Discussion

3.1. Reliability and validity of the rubric

Before analysing consistency and studying the differences among the various assessment methods, a study of the reliability and of the criterion-related validity of the rubric was conducted. Reliability assesses the internal consistency of the items, whereas criterion-related validity refers to their concurrent validity (Hair, Black, Babin, & Anderson, 2009). The instrument has a high reliability with a Cronbach's alpha of 0.850 for the students, and 0.886 for instructors. Regarding the second aspect, the correlation between the overall valuation item and the score obtained by adding up the other 3 items was examined. Positive correlations of 0.876 ($p < 0.001$) and 0.855 ($p < 0.001$) were obtained respectively for students and instructors, representing an acceptable criterion-related validity (Hair et al., 2009). Throughout the rest of the article, only the overall valuation item is considered.

3.2. Consistency among the different universities and assessment methods

Tables 1 and 2 show the consistency of the overall valuation item of products (N=36) between the universities (U1, U2) and assessment methods (self-, peer-, external- and instructor-assessment).

Table 1. Pearson correlation coefficients between assessments at both universities.

	U2 students	instructors
U1 students	0.843***	0.652***
U2 students		0.774***

***p<0.001

Table 2. Pearson correlation coefficients between the different types of assessments.

	Self	Peer	External
Instructor	0.443*	0.764***	0.772***
Self		0.700***	0.504**
Peer			0.863***

*p<0.05;**p<0.01;***p<0.001

A significant correlation can be observed between the assessments of the different participants. The effect size of the consistency is large (Ellis, 2010) in all cases (>0.5), except in the consistency between self and instructor assessment, which is medium (>0.3). If we consider in both tables the assessments of products from U2 and from U1 separately, similar correlations are observed. However, one aspect is slightly different. The correlation between self- and peer-assessment with U1 products is r=0.749 and with U2 products is r=0.465. When the products are considered together, the correlation

between U1 and U2 instructors is $r=0.867$ ($p<0.001$). If only U1 products (or U2 products) are examined, similar correlations coefficients are obtained.

These results support our first hypothesis. A high level of consistency exists between the different participants. It is worth mentioning that the consistency between peer- and external-assessment is the highest, and the consistency among instructor-, peer- and external-assessment is quite similar. The correlation between self- and instructor-assessment is the lowest.

Other researchers have reached conclusions similar to our findings. In the work presented in (Sadler & Good, 2006) and (Sung, Chang, Chiou, & Hou, 2005), the authors discovered a high correlation between self-, peer-, and teacher-assessment. A moderate correlation between teachers and peers was also found by (Panadero et al., 2013). In (Kulkarni et al., 2013; Luo et al., 2014), a high correlation factor between peer- and staff-assigned- grades in a MOOC course was also determined. Nevertheless, these findings are in conflict with other results. Consistency among peer-, self- and teacher-assessment was not observed by Chen (2010). The authors of (Chang et al., 2012) discovered the consistency between the results of self- and teacher-assessment with a high effect size (0.83), but they did not observe the consistency between self- and peer-assessment, or between peer- and teacher-assessment. In (Luo et al., 2014), the researchers obtained a low correlation factor between self- and staff-assigned grades in a MOOC course; and in (Hovardas et al., 2014), the usage of a single reviewer produced a low validity for the majority of peer assessors. The literature review presented in (Toppings, 2003) concludes that, in general, peer assessment seems likely to correlate more highly with instructor assessment, rather than self-assessment; and that self- and peer-assessment do not always correlate well. A recent meta-analysis (Li et al., 2015) confirms that peer-ratings generally exhibit a moderately high level of agreement with teacher-ratings. As explained in (Chang et al., 2012), the variety in the (obtained) results may be due to the varying educational levels of students, assessment rubrics, assessment environments, assessor trainings, number of reviewers for each product, and so on. For instance, the work presented in (Cho et al., 2006; Luo et al., 2014) suggests that multiple reviewers (a collection of three to six peers) should be used to ensure high consistency between peer- and instructor-assessment. In the present study, the means of the assessment used for each product is that proposed by each type of assessor (self, peer, external, and instructor). Moreover, the assessments did not have any specific requirement related to the subject or the academic level, and it is our belief that similar assessments could be conducted in other subjects (not specifically computer oriented) or with other types of students cohorts (not solely the university level). Our results are in line with the results obtained by the previously mentioned literature reviews (Toppings, 2003; Li et al., 2015) in terms of self, peer and instructor assessments. The correlations with external assessment represent an innovative finding for the literature. More research is required to confirm whether similar results are obtained in other subjects and academic levels.

3.3. Differences between the two universities' assessments

Table 3 provides the means (standard deviations) of the overall valuation item given by students and instructors regarding the selected products of U1 (N=18) and U2 (N=18).

In Table 3, the instructors' assessments are given altogether. The U1 instructors assessed the U1 and the U2 products with means (standard deviations) of 3.47 (0.81) and 4.1 (0.76) (t=-2.204, p=0.036), respectively; and the U2 instructors with means (standard deviations) of 3.21 (0.69) and 3.93 (0.71) (t=-2.823, p=0.009), respectively.

Table 3. Assessments given by students and instructors for selected products from both universities.

	U1 products	U2 products	Test (Student's t)
Instructors	3.31 (0.72)	3.99 (0.70)	t=-2.602, p=0.015 *
U2 students	2.99 (0.67)	3.88 (0.45)	t=-4.280, p=0.000 ***
U1 students	3.49 (0.56)	3.58 (0.38)	t=-0.486, p=0.631

*p<0.05, ***p<0.001

The overall perception is that the U2 products have better quality than the U1 products. Instructors (either grouped all together or separated by universities) and U2 students consider, with significant differences, that the products carried out at U2 are better than the ones at U1. It is remarkable that the U2 students assigned lower grades to U1 products. However, U1 students did not appear to notice such a difference and assigned lower grades to U2 products. There could be a number of reasons for this result. The U1 students may have been more generous with their own work, and they did not want to clearly state such a difference; or, it could be a cultural issue, i.e. students from one region may be less (or more) demanding than students from another region. Another alternative explanation could be competitiveness between the groups of students from different universities (i.e. students from one university might view students from the others to be rivals).

To address the aforementioned discrepancy, the products were classified only on the basis of instructors' assessments. Two groups of products were created: best products and worst products. The former group consists of the products whose assessments are equal or higher than the median; and, the latter group comprised the other products. The obtained medians were 3.8 for the U2 products, and 3.5 for the U1 products.

Table 4 gathers the results (means and standard deviations) from the overall valuation item of the U1 students, U2 students and instructors (considered altogether) regarding the best (N=20) and worst (N=16) products.

Table 4. Assessments given by students and teachers of best and worst products classified on the basis of instructor assessment.

	Worst products	Best products	Test (Student's t)
Instructors	3.01 (0.59)	4.13 (0.50)	t=-5.608, p=0.000 ***
U2 students	3.11 (0.81)	3.68 (0.56)	t=-2.277, p=0.031 **
U1 students	3.26 (0.45)	3.75 (0.38)	t=-3.196, p=0.004 **

**p<0.01, *** p<0.001

All the groups, including the U1 students, made a clear distinction between the best and worst products. In this way, students from both universities were equally demanding regarding the selected products; that is, the students from one region are not (culturally) more demanding than the students from the other region. However, it seems that students from one university are less demanding with the product of their university, while they are more demanding with the products of the other university.

Table 5 displays the student poll on the best product. The first row of the table gathers the data from all the students that expressed their preference. The last two rows split the data according to whether the surveyed student's product was assessed.

In Table 5, one can observe that, in general, most of the votes (134) were awarded to U2 products (71.7%) – as noted above, U2 products were generally considered to be better products by all the participants. However, it is also worth mentioning that 73.3% of the students chose products from their own university. If the teams that could assess themselves are excluded, 67.4% of the votes were given to peers from the students' same university. If only those votes from the students who could self-assess are considered, 78.9% selected products were from their own university. Moreover, the 31.1% (14 out of 54) of the U1 students, and 60.4% (29 out of 48) of U2 students chose their own product as the best product ($\chi^2 = 8.024$, $p=0.005$).

Table 5. Total amount of votes related to the best product according to students' opinion.

Students	U1→U1	U1→U2	U2→U2	U2→U1	Test (χ^2)
All	47	44	90	6	47.408***
Own product assessed	19	26	43	4	13.934***
Own product not assessed	28	18	47	2	35.412***

***p<0.001

These results support our second hypothesis: students from one university seem to be less demanding when assessing the products from their own university. Moreover, students tend to choose their products as the best products, and favour the products from their own university over the products from the other university.

In the (obtained) results, a group-influence can be observed: that is, people tend to favour members from their own group (Pronin, 2006), as opposed to outside people, who are in some way considered as competitors. In this line, the authors of (Kulkarni et al., 2013) observed that on average, students graded products from their own country higher than those from other countries. Cultural aspects are proposed as a possible reason for that "patriotic" bias because, in this research, grading was double-blind. In our case the students were aware of the origin of the products. These products came from universities located in nearby regions within the same country. Thus, it is reasonable to suppose that a cultural bias does not exist. Future research will have to determine this issue through a research design wherein the students are not aware of the origin of the products evaluated. The authors of (Sung et al., 2005) also asked for the best products at just one university. In that case, they took into account the best products in the first tertile, and they discovered an association between the best works selected by

self-, peer- and instructors. In our case, where only the best products from the two universities were to be selected, such an association was not observed. Finally, it is worth mentioning that the group influence did not occur among the instructors in our study; namely, both the U2 and U1 instructors chose U2 products as the best (85.7% and 87.5% respectively); there are not significant differences among them (χ^2=0.014, p=0.907).

3.4. Differences among the assessment methods

Table 6 shows the results (means and standard deviations) of the overall valuation item in the three kinds of assessments depending on student social-implication regarding the products from both universities. Table 6 also includes the instructors' assessments.

Significant differences can be observed when self-, peer- and external-assessments are compared. Self-assessments produce higher valuations than peer- and external-assessments in both universities. Significant differences can also be observed between peer- and external-assessments; however, these differences are lost after applying the Bonferroni correction in the U2 products. Regarding the grades awarded by the instructors, there are not significant differences with peer-assessments, but there are differences when compared to self- and external-assessment. The former are eliminated after applying the Bonferroni correction in the case of U2.

Table 6. Means (standard deviations) of student assessments of products from both universities depending on their social implication with the development team. Instructor assessment has also been included.

Products	Self (S)	Peer (P)	External (E)	Test (rANOVA)	After Bonferroni	Instructor (I)	After Bonferroni
All	4.16 (0.57)	3.59 (0.54)	3.28 (0.61)	F=55.685***	S>P>E	3.65 (0.78)	S>I≃P>E
U1	4.15 (0.49)	3.36 (0.55)	2.99 (0.67)	F=90.323***	S>P>E	3.31 (0.72)	S>I≃P>E
U2	4.17 (0.67)	3.82 (0.43)	3.58 (0.38)	F=11.206***	S>P≃E	3.99 (0.70)	I≃P, I≃S>E

***p<0.001; > there are significant differences, ≃: there are not significant differences.

In spite of the fact that the U2 products seem to be better than the U1 products, self-assessments were quite similar at both universities (t=-0.083, p=0.934). However, peers (t=-2.583, p=0.015) and external-peers (t=-2.973, p=0.006) reveal these differences in their assessments.

These results support our third hypothesis: students award themselves better grades than peers do, and external-peers are even stricter.

Now let us examine the differences among the three kinds of assessments depending on student social implication and regarding the best and worst products (according to the instructors' criteria). Table 7 displays these results (means and standard deviations) and also includes the valuations of the instructors.

Table 7. Means (and standard deviations) of student grades given to the best and worst products (according to the instructors' criteria). Student social implication is noted, and instructor assessments is also included.

Products	Self (S)	Peer (P)	External (E)	Test (rANOVA)	After Bonferroni	Instructor (I)	After Bonferroni
Worst	3.97 (0.49)	3.31 (0.54)	2.94 (0.67)	F=26.842***	S>P>E	3.01 (0.59)	S>P>I≃E
Best	4.30 (0.61)	3.80 (0.44)	3.54 (0.42)	F=29.446***	S>P>E	4.13 (0.50)	S≃I>P>E

***p<0.001; > there are significant differences, ≃: there are not significant differences.

The results in Table 7 show that the differences among self-, peer- and external-assessment are maintained. Hence, these differences are independent of product quality. In addition, even if the mean self-assessment of the best products is higher than that of the worst products, the difference is not significant (t=-1.614, p=0.118). Nevertheless, in the peer- and external-assessments, there are significant differences depending on the quality of the product (t=-2.735, p=0.011 in the former, and t=-2.995, p=0.006 in the latter). Finally, instructor- and external-assessments are similar for the worst products; and, instructor- and self-assessment are similar for best products.

In general, when students assess themselves, they are more generous than when they assess their peers. Moreover, self-assessments do not reveal the distinction between best and worst products. In this way, self-assessment is similar to instructor-assessment for the best products (although, it is higher), but is not at all similar for the worst products. It can also be assumed that students tend to homogenise assessment when acting as peer assessors, in that they award a slightly lower grade than the instructor's grade to the best products, and a slightly higher grade to the worst products, although they do distinguish between good and bad products. In the case of external peers, they are by far the strictest assessors. Such strictness is similar to that of instructors in the case of worst products, but is not at all similar in the case of best products. This finding supports the first part of our fourth hypothesis: when considering the mean of the given assessments, instructor-assessment is the closest to peer-assessment, but depending on the quality of the product it may be closer to self- or external-assessment.

More generous self-assessment has previously been documented in the literature, see for instance (Sadler & Good, 2006; Chang et al., 2012); but the more demanding external-peer-assessment is a new observation. This strictness can be interpreted in a variety of ways: formative (students from one university are not able to understand others' work habits), rivalry or group-identity reasons – students may believe that students from other universities represent future rivals (Pronin, 2006). In the case of instructors, this effect does not appear, as noted in Section 3.1, teachers from both universities assigned better grades to the U2 products. Interestingly, the U1 instructors awarded higher grades, with a mean of 4.1 (0.76), to the U2 products as opposed to the U2 instructors, mean of 3.93 (0.71). For future research, it would be interesting to introduce a double-blind review in the assessment process (assessors would not be aware of whether or not the assessed product belongs to their peers). Such research could be compared with the results

presented in (Snodgrass, 2007), where single-blind versus double-blind reviewing is analysed.

The impact of friendship on peer-assessment has been previously studied in the literature. In (Panadero et al., 2013), the degree of friendship among students is taken into account. Though it is obtained that, in general, peers tend to over-score, there is even more noticeable over-scoring by students with a stronger degree of friendship. However, the study presented in (Azarnoosh, 2013) revealed no significant difference between ratings of friend- and non-friend-peers. The latter study mentions that the difference in findings may be due to the general familiarity and friendship of all the students with one another in the class. Although students identified their closest friends, they did not deny their overall friendship with others who had been their classmates for at least 2 years, so this may have affected their ratings subconsciously. Another issue is the possible fear of facing those friends the following week in class after giving someone a bad grade. In our study, there is not a friendship relation with external-peers, and the friendship level among students of the same university was not considered. However, the results indicated that the grades given by peers that did not know each other were stricter than the grades awarded by peers that have studied together for several years.

There is heterogeneity in the literature regarding the results when comparing peer- and teacher-assessment. In (Sadler & Good, 2006; Pereira et al., 2014), peers awarded lower grades than teachers. On the contrary, the results obtained in (Chang et al., 2012), based on the assessment of a single project (the design and implementation of a website), showed that teachers are stricter than peers. And lastly, no significant difference was found in (Azarnoosh, 2013) between teacher- and peer-ratings of students' English compositions. Similarly, the authors of (Cho et al., 2006) also found that student ratings are as valid as instructor ratings of English writing. The authors of (Hamer, Purchase, Luxton-Reilly, & Denny, 2015) also discovered that there was no significant difference between grades of peers and instructors in a large, undergraduate software engineering programming class. From our point of view, these heterogeneous results are due to a lack of distinction between the quality of products; an issue that was tackled in the present study. A similar distinction was made in (Sadler & Good, 2006); namely, the authors obtained that, when grading others, students awarded the best-performing students lower grades than their instructors did.

3.5. Assessment awarded according to student competency

Table 8 displays the results (means and standard deviations) of the overall valuation item given by 3 groups of students. The groups were created based on students' final grade (on a scale from 0 to 10) for the course: the highest-grade group (students with a grade higher than 8, N=31), the intermediate-grade group (students with a grade higher than 7 but lower than 8, N=32), and the lowest-grade group (students with a grade lower

than 7, N=34). The assessments given by these three groups were compared in terms of self-, peer, and external-assessments.

In the previous sections, the assessments of the products considered altogether were compared, or divided into different groups based on either the origin of the product or its quality. However, now the products are divided into groups of different assessors. When comparing these groups, the same group of products are not under consideration and, therefore, a normalisation step is required. Namely, the mean grade of instructors has been subtracted from the grade given by the students. This method has been used in the literature; for instance, in (Sadler & Good, 2006): they considered this difference as the error in student assigned-grades.

In Table 8, a similar behaviour can be observed for the three groups of students. Students gave themselves better grades than their peers, and they are more demanding of external-peers. Only on the case of the highest-grade group, did students assess themselves similarly to their peers. There are significant differences between the grades given among peers from the same university and externals in the cases of highest-grade and lowest-grade groups, but these differences disappear after applying the Bonferroni correction.

Table 8. Assessments of different groups (student groups divided according to their final grade).

Students	Self (S)	Peer (P)	External (E)	Test (rANOVA)	After Bonferroni
Highest-grade	-0.02 (0.92)	-0.18 (0.59)	-0.48 (0.43)	F=3.621*	S>E, P≃S, P≃E
Intermediate-grade	0.52 (0.71)	-0.01 (0.36)	-0.46 (0.42)	F=31.871***	S>P>E
Lowest-grade	0.81 (1.22)	-0.19 (0.44)	-0.48 (0.41)	F=24.438***	S>P≃E

*p<0.05, ***p<0.001; > there are significant differences, ≃: there are not significant differences.

The most interesting fact can be observed when self-assessments of the three groups are compared: in this situation, significant differences are observed ($F=4.182$, $p<0.05$). The students of the lowest-grade group tended to give themselves better grades than the rest – students with the lowest grades usually produce lower-quality products, and they were probably trying to compensate for this fact. Moreover, students of the highest-grade group tended to assess their work similarly to the instructor. However, there are not significant differences between the levels of expectation of peers and external peers ($F=1.108$, $p=0.336$; $F=0.019$, $p=0.981$, respectively) for any of the three groups.

These results partially support Part (ii) of our fourth hypothesis: all students similarly assess products from peers and external peers regardless of their final grade. However, students self-assess differently as they attempt to place their product's quality in the highest category.

The Dunning-Kruger effect (Kruger & Dunning, 1999) could explain this situation. According to these authors, unskilled individuals tend to overestimate their abilities in many social and intellectual domains. Their incompetence robs them of the metacognitive ability to realise their distorted perceptions. Moreover, paradoxically, by

improving their skills, individuals can better recognise the limitations of their own abilities. Results analogous to ours were obtained by (Schlösser, Dunning, Johnson, & Kruger, 2013); in this research, students tended to overrate their performances (poor performers produced self-assessments similar to the ones produced by those performing at the top); additionally, the ability of the top-performers to accurate evaluate themselves was also observed. Similar results were also obtained by (Topping, 2003) and (Sadler & Good, 2006). In the former study, more capable students tended to under-rate themselves, while weaker students over-rated themselves to a larger greater extent. In the latter study, poorly performing students tended to over-rate themselves in comparison to teacher-assigned grades.

3.6. Student learning

Table 9 shows a comparison of the grades (means and standard deviations) of the students from both years when the student assessments were collected (2013 and 2014), and from the year before (2012). The scale used to measure the grades ranges from 0 to 10. In the academic year 2012, the same teaching-methodology was employed, but student assessments were excluded (self, peer, and external). Lilliefors-corrected Kolmogorov-Smirnov test was applied to check whether the grade variable for all students (and for the U1 and U2 students) with assessments followed a normal distribution. This hypothesis can be rejected since the test was =0.111, p=0.005 (=0.192, p<0.001 for U1, and =0.150, p=0.007 for U2). The non-parametric Mann-Whitney U test is included here.

Table 9. Grades (means and standard deviations) obtained by the students during two years (2013 and 2014, N=97) when this study was conducted and the year before (2012, N=25).

Students	without assessments	with assessments	Test (Mann-Whitney U)
All	6.76 (0.92)	7.20 (1.34)	Z=-1.974; p=0.048*
U1	6.82 (0.91)	7.24 (1.20)	Z=-2.148; p=0.032*
U2	6.6 (1.0)	7.16 (1.49)	Z=-1.116; p=0.264

In the results, a significant improvement can be observed in the final grade obtained by students who participated in the assessment experience as compared to students from the previous year. These differences are maintained in U1, but not in U2. Given the small number of U2 students (N=7) that participated in the year prior to the assessment experience, the size effect is taken into consideration, obtaining a d of Cohen value of 0.44, which is close to a medium effect (Ellis, 2010).

These results confirm our fifth hypothesis: the assessment experience may translate into an improvement in the students' final grades as compared to the previous course when the same learning method was implemented, but excluding student assessment.

The implications of this research in terms of student learning further support the well-documented results in favour of the formative influence of peer-assessment. This result

has been described in many previous studies (see, for instance, (Chen et al., 2009) at traditional universities, or (Kulkarni et al., 2013) at distance-learning universities, such as in a MOOC course). However, future experiments ought to contrast whether there are differences among the motivation and formation that can arise from the rivalry produced between external-assessment and peer-assessment.

4. Implications and Conclusions

This article presents a comparison among online self-, peer-, external-, and instructor-assessments of web products in two similar courses from two different universities. First of all, a high-level consistency was observed across the different kinds of participants. In addition, a competitive effect was discovered: students from one university seemed to be less demanding when assessing products from their own university, whereas they tended to choose their own products as the best, as well as the products from their own university in favour of products from the other university. And lastly, it was observed that students awarded themselves better grades than their peers did; and external-peers were even stricter. This effect may be due to the competitive nature of groups. Although, in this study two different groups were clearly defined (one from each university), different types of groups can arise both in traditional educational settings and in MOOC courses: for instance, based on a common native language. Therefore, this type of possibility should be taken into account when conducting peer-assessment. Moreover, the type of assessments developed in this study did not have any specific requirement related to the subject or the academic level. It is our belief that this competitive effect could also occur in other subjects (not necessarily computer-oriented) and with other type of students cohorts (not specifically at the university level), but additional research is necessary to prove to confirm this.

And secondly, this study has identified differences in the implemented assessments depending on the quality of the assessed-products and the competency of the assessing student. In the case of the former factor, variations were observed: for low-quality products, external-assessments are the closest to instructor-assessments; for high-quality products, self-assessments are the most similar to instructor-assessments; and, on average, peer-assessments are the closest to instructor-assessments. Regarding the competency of the assessing student, students (regardless of the final grade they obtained for the course) assess products from peers and external peers in the same way; however, they self-assess differently: aiming to place the quality of their products at the same level of the best products. In this way, it should be taken into account that, although peer assessment does not seem to depend on student competency, it does appear to depend on product quality.

Finally, our study’s results seem to indicate that students’ motivation was increased by observing products made by their peer and external students and by the competitiveness that this comparison creates. This translates into an improvement in the students’ final grades as compared to the previous year when peer assessment was not included. In this

line of thought, the question remains as to whether the external-assessment is necessary to improve the students results, or if peer-assessment suffices.

References

Azarnoosh, M. (2013). Peer assessment in an EFL context: attitudes and friendship bias. *Language Testing in Asia*, 3(11).

Boubouka, M., & Papanikolaou, K. A. (2013). Alternative assessment methods in technology enhanced project-based learning. *International Journal of Learning Technology*, 8(3), 263-296.

Chang, C.-C.,Tseng, K.-H., & 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, 303–320.

Chen N.-S., Wei, C.-W., Wua, K.-T., & Uden, L. (2009). Effects of high level prompts and peer assessment on online learners' reflection levels. *Computers & Education*, 52, 283–291.

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, 98, 891–901

Domínguez, C., & Jaime, A. (2010). Database design learning: A project-based approach organized through a course management system. *Computers & Education*, 55(3).

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes Statistical Power, Meta-Analysis, and the Interpretation of Research Results textbook*. Cambridge University Press.

Falchikov, N. (2003). Involving students in assessment. *Psychology Learning and Teaching*, 3(2), 102–108.

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.

Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2009). *Multivariate Data Analysis*. 7th ed. Pearson.

Gielen S., Dochy F., & Onghena P. (2011). An inventory of peer assessment diversity. *Assessment and Evaluation in Higher Education*, 36, 137–155.

Hamer, J., Purchase, H., Luxton-Reilly, A., & Denny, P. (2015). A comparison of peer and tutor feedback. *Assessment & Evaluation in Higher Education*, 40(1), 151–164.

Hovardas, T., Tsivitanidou, O. E., & Zacharia, Z. C. (2014). Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. *Computers & Education*, 71, 133–152.

Jaime, A., Domínguez, C., Sánchez, A., & Blanco, J. M. (2013). Interuniversity telecollaboration to improve academic results and identify preferred communication tools. *Computers & Education*, 64, 63–69.

Kennedy, G. J. (2005). Peer-assessment in group projects: Is it worth it? Proceedings of the 7th Australasian Conference on Computing Education, 42, 59–65.

Kizilcec R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. Proceedings of the Third International Conference on Learning Analytics and Knowledge. ACM, 170–179.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134.

Kulkarni, C., Koh, P. W., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., & Klemmer, S. R. (2013). Peer and self-assessment in massive online classes. *ACM Transactions on Computer-Human Interactions*, 9(4), 39.

Lai, C.-L., & Hwang, G.-J. (2015). An interactive peer-assessment criteria development approach to improving students' art design performance using handheld devices. *Computers & Education*, 85, 149–159.

Legault, L., & Amiot, C. (2014). The role of autonomy in intergroup processes: Integrating Self-Determination Theory and intergroup approaches. In N. Weinstein (Ed.), Integrating Human Motivation and Interpersonal Relationships: Theory, Research and Applications, 159–190. Springer.

Llado, A. P., Soley, L. F., Sansbello, R. M. F., Pujolras, G. A., Planella, J. P., Roura-Pascual, N., et al. (2014). Student perceptions of peer assessment: an interdisciplinary study. *Assessment & Evaluation in Higher Education*, 39(5), 592–610.

Li, L., Liu, X., & Steckelberg, A. L. (2010). Assessor or assessee: How student learning improves by giving and receiving peer feedback. *British Journal of Educational Technology*, 41(3), 525–536.

Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., Lyu, Y. Chung, K. S., & Suen H. K. (2015). Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education*, DOI:10.1080/02602938.2014.999746.

Lin, S. S. J., Liu, E. Z. & Yuan, S. M. (2001). Web-based peer assessment: feedback for students with various thinking-styles. *Journal of Computer Assisted Learning*, 17, 420–432.

Luo H., Robinson A. C., & Park J. Y. (2014). Peer grading in a MOOC: Reliability, validity, and perceived effects. *Online Learning Journal*, 18(2), 1–14.

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.

Panadero, E., Romero, M., & Strijbos, J. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation*, 39(4), 195–203.

Pereira, J., Echeazarra, L., Sanz-Santamaría, S., & Gutiérrez, J. (2014). Student-generated online videos to develop cross-curricular and curricular competencies in Nursing Studies. *Computers in Human Behavior,* 31, 580–590.

Pronin, E. (2006). Perception and misperception of bias in human judgment. *Trends in Cognitive Sciences*, 11(1), 37–43.

Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment*, 11(1), 1–31.

Schlösser, T., Dunning, D., Johnson, K. L., & Kruger, J. (2013). How unaware are the unskilled? Empirical tests of the ‘‘signal extraction’’ counterexplanation for the Dunning–Kruger effect in self-evaluation of performance. *Journal of Economic Psychology*, 39, 85–100.

Snodgrass, R. T. (2007). Editorial: Single-versus double-blind reviewing. *ACM Transactions on Database Systems*, 32(1), 1.

Søndergaard, H., & Mulder, R. A. (2012) Collaborative learning through formative peer review: pedagogy, programs and potential. *Computer Science Education*, 22:4, 343-367.

Strijbos, J. W., & Sluijsmans, D. (2010). Unravelling peer assessment: Methodological, functional, and conceptual developments. *Learning and Instruction*, 20, 265-269.

Sung, Y.-T., Chang, K.-E., Chiou, S.-K., & Hou, H.-T. (2005). The design and application of a web-based self and peer-assessment system. *Computers & Education*, 45, 187–202.

Topping, K. J. (1998). Peer Assessment between Students in Colleges and Universities. *Review of Educational Research*, 68(3), 249–276.

Topping, K. J. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In Optimising New Modes of Assessment: In Search of Qualities and Standards, vol. 1, 55–87. Springer.

Topping, K. J. (2010). Methodological quandaries in studying process and outcomes in peer assessment. *Learning and Instruction*, 20(4), 339–343.

Tseng, S. C., & Tsai, C. C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education*, 49(4), 1161–1174.

Tynjälä, P., Pirhonen, M., Vartiainen, T., & Helle, L. (2009). Educating IT project managers through project-based learning: A working-life perspective. *Communications of the Association for Information Systems*, 24(1), 16.

Willmot, P., Pond, K., Loddington, S. P., & Palermo, O. A. (2008). Perceptions of peer assessment in university teamwork. In International Conference on Engineering Education, 27–31.

Appendix

Table A.1. Description of the method applied based on clusters proposed by Gielen et al., 2011.

Decisions concerning the use of peer assessment	
Setting	Students from two distant universities: U2 and U1. Courses about computing project management in the third level of a computing engineering degree program. Total number of participants: 97.
Object	Products created in teamwork: simple multimedia web system.
Frequency-experience	Three valuation cycles.
Objectives	Combination of learning and learning-how-to-assess. Students become aware of quality definition, assurance and control.
Function	Formative.
Link between peer assessment and other elements in the learning environment	
Alignment	Competences in computing project management: handle concepts related to engagement with the quality in professional environments. The student must look for the consistency between his or her evaluation and that of his or her peers.
Relationship to other assessments	Students assess students from the same university and from another university. Expert instructors from both universities also assess. Selected products are also self-assessed. None of the valuations influence the final grade (neither self-assessment nor peer-assessment).
Scope of involvement	The students develop a product of the same characteristics as the assessed products. In the same session, a set of products is assessed: assessment is conducted at the same time as the comparison. The student assesses according his/her own criteria without knowing the instructor's opinion.
Interaction between peers	
Output	Simple web forms about basic aspects. Mainly quantitative aspects. Ranking (the best product is selected).
Directionality	Mutual. All students assess the same set of products.
Privacy	Assessments are not anonymous for the instructors, but they are for the assessed students.
Contact	The form is completed online independently, and in an asynchronous way prior to a deadline.
Role of assessors	Passive. Reflexive effect, students are aware of how their work is perceived by others.
Composition of assessment groups	
Matching	The instructor selects the teams to create products. Different teams are created for each new product.
Assessors and assesses	Assessments are assigned individually. The assessed product has been created by a team. A selection of products is assessed. The selected products are assessed by all the participants.
Management of assessment procedure	
Format	Assessment format established by instructors.
Requirement	Assessment is compulsory for all students.
Reward	None
Training/guidance	None.
Quality control	Strongly dissonant, without contrast, partially performed assessments are not allowed.