Peer-assessment in higher education – twenty-first century practices, challenges and the way forward

Article *in* Assessment & Evaluation in Higher Education · October 2015

DOI: 10.1080/02602938.2015.1100711

CITATIONS
22

READS
778

1 author:

Michael Mogessie
Carnegie Mellon University
10 PUBLICATIONS **65** CITATIONS

SEE PROFILE

Peer-Assessment in Higher Education – 21st Century Practices, Challenges, and the Way Forward

Michael Mogessie Ashenafi[a,*]

[a] *Department of Information Engineering and Computer Science, University of Trento, via Sommarive 5, 38123 Trento, Italy*

* Corresponding author. Tel.: +393932689361; fax: +39 0461 28 2093 E-mail address: michael.mogessie@unitn.it

Own publications cited in this article:

Ashenafi, M. M., Riccardi, G., & Ronchetti, M. (in press). Predicting Students' Final Exam Scores from their Course Activities. In *Frontiers in Education, 2015*. IEEE.

Peer-Assessment in Higher Education – 21st Century Practices, Challenges, and the Way Forward

Peer-assessment in higher education has been studied for decades. Despite the substantial amount of research carried out in the discipline, peer-assessment has yet to make significant advances. This review identifies themes of recent research in peer-assessment practices and highlights the challenges that have hampered its advance. Most of these challenges arise from the manual nature of peer-assessment practices, which prove intractable as the number of students involved increases. Practitioners of the discipline are urged to forge affiliations with closely related fields and other disciplines such as computer science in order to overcome these challenges.

Keywords: peer-assessment; formative assessment; summative assessment; higher education

Introduction

Educational assessment can have formative or summative goals. Summative assessment is intended to measure the extent to which a student has achieved pre-specified learning goals. This type of assessment is commonly carried out at certain intervals throughout a course (Harlen & James, 1997; Morgan & O'reilly, 1999; Myers 2008). Criterion-referenced summative assessment measures the achievement of a student against clearly stated public standards regardless of the performance of other students in the class, whereas norm-referenced summative assessment evaluates the performance of a student against standards that are set according to the achievements of all students in the group (Harlen & James, 1997; Morgan & O'reilly, 1999).

Formative assessment is rather student-centred – it is intended to provide continuous support and feedback to students in order that they monitor their own progress and identify their strengths and weaknesses. It also helps teachers adjust their instruction in accordance with the progress of the class. Formative assessment is commonly intended to bear no summative value – it should not contribute towards final grades and students should be kept informed on their results (Harlen & James, 1997; Morgan & O'reilly, 1999; Myers, 2008).

Several teaching-learning environments have adopted formative assessment to take advantage of its intended benefits. Non-traditional environments, where the teacher is not the sole assessor of a student's work, commonly use either pure formative assessment or a blend of formative and summative assessment. In these environments, students are heavily involved in evaluating their works as well as those of their peers.

In peer-assessment, students or groups of students assess the works of other students, their peers. Topping (1998) defines peer-assessment more formally as "… an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status" (p. 250).

A significant amount of research in the area of peer assessment, with focus on a wide array of themes, has been conducted since the turn of the century. The purpose of this study is to provide a comprehensive review of research in peer-assessment conducted during this

period, to highlight the challenges practitioners face, and to recommend ways to overcome these challenges in the quest to revolutionise educational assessment.

Peer Assessment in the 21st Century

An extensive review of literature of the past century regarding peer-assessment by Topping (1998) revealed factors that varied throughout the 109 studies that were reviewed. Among these factors were the wide variation in curriculum areas or subjects, the objectives of peer-assessment projects, whether peer-assessment was conducted in a formative or summative manner, variation in the work being assessed, and varying degrees of agreement between peer- and teacher-assigned scores.

Topping concluded that, given the varying nature of these factors throughout the studies considered, it was difficult to make concrete conclusions about the soundness or practicality of peer-assessment in higher education courses, or to provide a general theoretical model for peer-assessment.

Falchikov and Goldfinch (2000) conducted a meta-analytic review of studies that compared peer- and teacher assigned marks. They identified population characteristics, the work being assessed, the course level, the nature of assessment criteria, and the number of teachers and students involved per assessment task as the variables that affected the quality of the studies.

Bangert-Drowns, Wells-Parker, and Chevillard (1997) outline criteria typically used by meta-analytic reviewers when assessing the quality of research, present statistical strategies that help define the notion of study quality, and discuss how these criteria and strategies can be used to make reliable judgements of study quality. Falchikov and Goldfinch subsequently applied these study quality assessment measures when evaluating the quality of the experimental design of the studies.

Falchikov and Goldfinch summarised their study by concluding that, on average, peer marks agreed with teacher marks. They identified six factors that were most likely to influence improvements in agreement between teacher and peer assessments. The factors identified were:

- Asking peers to provide overall judgements based on well-specified assessment criteria
- Peer assessment in educational environments seemed more effective than in professional settings
- Better experimental designs led to better peer-teacher assessment agreements
- While there was no indication that multiple peer ratings increased agreement between peer and teacher agreements, increase in the number of peers evaluating a single work seemed to lower agreement scores
- While there was no indication in the validity of peer assessment regarding subject areas, there were some cases where student assessments in medical subject areas tended to agree less with those of teachers'
- Student-defined and well-understood criteria tended to lead to better agreements between peer and teacher assessments.

Inclusion Factors

The keywords *peer assessment*, *peer grading*, *peer evaluation*, *peer review*, *peer feedback*, and *peer interaction* were used to search for relevant literature. The search was carried out on Google Scholar[1].

After a brief analysis of the contents of the studies returned in the search results, all studies that did not discuss any aspect of peer-assessment in detail, that were not conducted in a higher education setting, or that involved peer-assessment of professional practice rather than academic products and processes were subsequently excluded from the analysis. Because this study is intended to extend the reviews discussed in section 2 (Topping, 1998; Falchikov & Goldfinch, 2000), all studies dated before 2000 were also excluded. A further list of studies that were cited by the studies that were returned by the search and that met the inclusion criteria was added to the initial list.

Fourteen studies discussing computer-based or web-based peer-assessment tools were identified in either the search results or in the citations. Because all those studies were published before 2009, the reader is encouraged to see Luxton-Reilly (2009) for a comprehensive review of the predominant computer-based, web-based, or electronic peer-assessment platforms in use by institutions of higher education today.

Authors were contacted in order to obtain copies of studies that were not freely available on the Internet. The final list of papers reviewed in this study is comprised of 64 articles, published in several journals and conference proceedings.

Literature Reviews

The discussion of the following review papers helps identify areas that have been the focus of peer-assessment research in recent years and where the main problems in peer-assessment processes reside.

Kollar and Fischer (2010) note that peer-assessment is still in its infancy despite decades of research in the field. They stress that it needs to establish affiliations to closely related practices such as collaborative learning. This view is also shared by Strijbos and Sluijsmans (2010) who argue that opportunities in advances in similar fields have not been taken advantage of.

The subsections that follow discuss similar categories of issues in peer-assessment that have been identified by scholars and practitioners of peer-assessment.

Student Involvement

Several studies recommend that students be actively involved in the various stages of peer-assessment. Falchikov (2003) argues that any assessment task must have students as active participants in order for it to be effective. Falchikov states that any peer-assessment should allow replication and students must be given clear instructions regarding the processes of the peer assessment.

The importance of student involvement in all stages of peer-assessment is also highlighted by Tillema, Leenknecht, and Segers (2011). The importance of involving students in the specification of assessment criteria has also been stressed by other studies (Sluijsmans, Brand-Gruwel, van Merriënboer, & Martens, 2004; Bloxham & West, 2004).

[1] Google Scholar – http://scholar.google.com

The Variables of Peer Assessment

A number of studies have determined important variables that are common in many peer-assessment practices.

Psychometric qualities, domain-specific and peer-assessment skills, and students' attitudes towards peer-assessment are four variable categories Van Zundert, Sluijsmans, and Merriënboer (2010) investigate in their review of a selection of 26 articles published between 1990 and 2007.

Topping (2010) identifies some uncertainties in peer-assessment and argues that it should be explored in detail whether peer-peer relationships have an impact on the process, whether peer-feedback should be iterative or a one-off process, and if assigning multiple peers to the same assessment task is more effective. In his review, Topping reveals inconsistencies or contradictory results and flaws or limitations in experimental designs of the studies concerned.

van den Berg, Admiraal, and Pilot (2006a) select 10 of the 17 variables identified by Topping (1998) that they consider important for an optimal peer-assessment design. They identify as important features the type of product being assessed, whether the assessment is a substitution for staff assessment, whether it is mutual and anonymous, whether assessor-assessed contact is face-to-face, whether the abilities of group members are equivalent, whether peer-assessment is individual or group-based both for assessor and assessed, whether peer-assessment is in-class or out-of-class, and whether there is a reward for participating in peer-assessment tasks.

In a later study, the authors experiment with varying degrees of these variables to determine their impact on oral and written feedback (van den Berg, Admiraal & Pilot, 2006b). They conclude that in order for peer feedback to be optimal, peer-assessment should be done in small groups, with either formative or summative goals, and that written feedback should be orally explained and discussed with the assessed. This, however, raises the issue of whether such peer-assessment is practical in large classes.

Interpersonal variables are also considered to affect learning outcomes in peer-assessment. van Gennip, Segers, and Tillema (2009) identify psychological safety, value diversity, interdependence, and trust as four interpersonal variables that have an impact on learning outcomes in peer-assessment. Except task interdependence, which refers to the responsible involvement of students in peer-assessment tasks, the interpersonal variables identified by van Gennip, et al. (2009) are specific to group-based peer-assessment tasks.

Quality of Peer Assessment

Reviews of literature in peer-assessment show that, although peer-assessment quality is often discussed by many studies, there have been few studies that evaluate the quality of their peer-assessment methods and that quality standards in peer-assessment have yet to be formally defined as practitioners rather set out their own quality measurement criteria.

Tillema, Leenknecht, and Segers (2011) review literature that discusses and applies measurements of quality in peer-assessment and outline three quality criteria that should be met at all stages of the peer-assessment process. These are:

Authenticity – often relates to actively engaging students in the assessment process to maintain relevance. It is linked in the literature to four specific criteria – representativeness, meaningfulness, cognitive complexity, and content coverage.

Transparency – refers to the quality of assessment tasks to be clear, understandable, and doable by those being assessed

Generalisability – refers to the extent to which the outcome of an assessment task can be generalised to a broader set or related tasks that measure the same achievement. It is linked in

the literature to four specific criteria – comparability, reproducibility, transferability, and educational consequences.

This contrasts with the views of Gielen, Dochy, Onghena, Struyven, and Smeets (2011), who demonstrate that the quality criteria that should be met in peer-assessment are determined by the goal of the peer-assessment task. This approach of specifying quality criteria is perhaps more practical as peer-assessment is implemented in such a wide variety of contexts that not many peer-assessment practices would meet a single set of quality measurement criteria.

Case studies, Action Research, Peer-Assessment Instruments

The studies in this category investigate specific settings in peer-assessment by conducting experiments that intend to measure relationships between variables.

The Value of Peer Feedback

The specificity of peer-assessment criteria has been shown to affect the quality of peer feedback – more specific criteria tend to provide more discriminative power to the assessment task at the risk of diminishing the quality of peer feedback (Miller, 2003).

A study regarding the nature and impact of peer feedback as perceived by 89 graduate students has shown that elaborate and specific feedback from peers, although found adequate, was perceived as having negative impact by the students receiving such feedback (Strijbos, Narciss & Dünnebier, 2010). The study states that the degrees of specificity and brevity have varying impacts on students with different levels of competence.

The impact of feedback on those who provide it has also been explored. Lin, Liu, and Yuan (2001) report on the specificity of peer-feedback and how students with various ways of thinking react to it and suggest specific feedback is more helpful than holistic feedback in improving students' performance.

It is also claimed that those students that provide their peers with high quality feedback tend to incorporate feedback from their peers effectively, raising their final grades in the process (Althauser & Darnall, 2001; Tsai, Lin, & Yuan, 2002). While another study could not confirm the existence of such relationship, it found a significant relationship between the quality of feedback a student provided and the quality of their own final project (Li, Liu, & Steckelberg, 2010). How the nature of peer-feedback and the number of peers providing it influence revision of initial work by the receiver has also been studied by Cho and MacArthur (2010), who suggest students receiving feedback from multiple peers tend to perform complex revisions of their work and produce higher quality products.

Other studies have stressed that training students on providing feedback, and in peer-assessment skills in general, improves the quality of feedback and as a result the quality of the final version of the product being assessed (Saito, 2008; Min, 2006; Hu, 2005; Sluijsmans & Prins, 2006). According to Chen and Tsai (2009), however, the improvement in quality becomes less significant in subsequent sessions of peer-review.

Peer Assessment Design Strategies

In a class of 12 students enrolled in a two-year postgraduate course, Topping, Smith, Swanson, and Elliot (2000) conducted a formative feedback-based peer-assessment experiment in which each student reviewed an academic report of their peer submitted at the end of the second term of the first year of the programme.

Although participation was mandatory, assessment results did not contribute towards final marks. Assessment tasks were completed out of class and anonymity of both parties was maintained. Assessment was reciprocal and paired.

Both staff and students used 14 criteria when assessing the academic reports. For each criterion, assessors provided a positive, neutral, or negative rating for the academic report. Only one student and one member of staff assessed each report, except in cases of possible fails, when double or triple staff assessments were carried out.

The study by Topping, Smith, Swanson, and Elliot sought to investigate agreement between staff and student ratings of the academic reports. Consequently, the total number of positive, neutral, and negative flags as well as the mean and standard deviation of the flags for student and staff ratings were computed. The authors reported percentages of overall positive and negative flags and overall positivity – the difference between positive and negative flags. The authors then concluded that there was an overlap in detail between student and staff assessments and that the validity and reliability of the approach appeared adequate, while admitting the finding may not generalise to other peer-assessment settings.

As shall be witnessed subsequent sections of this study, the study by Topping, Smith, Swanson, and Elliot demonstrates what many peer-assessment practices share and yet how dissimilar they are in the sense that findings of one study support or contradict those of another, if they do at all. A relatively small number of students, one-off experiments, incomparable results, and 2 or fewer assessors per task are common to many poorly designed peer-assessment practices while those that exhibit high quality design usually maintain anonymity, apply pre-specified assessment criteria, involve significantly large number of students, use multiple students per assessment task, and are conducted repeatedly among a group of students. The study by Topping, Smith, Swanson, and Elliot preserves anonymity and uses well-designed criteria but its findings are indeed not generalisable as it involves only 12 students and uses percentages to report teacher-student score agreements.

Ballantyne, Hughes, and Mylonas (2002) conducted a three-phase study spanning a two-year period involving 1654 students and 30 staff from three departments. Peer-assessment procedures outlined in the initial phase were revised together with students and faculty and re-implemented in subsequent phases.

Despite its high quality, the study by Ballantyne, Hughes, and Mylonas, which utilises an action research process in the design of peer assessment procedures, lacked qualities that would promote sustained implementation of the proposed approach. The distribution of assignments to peers was still manual. Moreover, given the high number of students involved, such was the effort needed to implement anonymous peer-assessment that some departments subsequently opted to forgo anonymity.

Depending on how it is implemented, peer-assessment may imply an increase or decrease of assessment-related load on teachers. In this particular case, the increase in load was shifted to students as students were required to meet outside class every week in order to exchange assignments and agree on final grades. Moreover, lack of anonymity increased the risk of bias.

Using automated peer-assessment tools, teachers could afford to enjoy the advantages that come with peer-assessment without the negative impacts discussed here because such tools offer anonymity and can easily automate assignment distribution, discussions, and submission of feedback and grades. Automated assessment can also help with calibrating grades assigned by multiple peers (Hamer, Ma, & Kwong, 2005).

The most common implementation of peer-assessment in higher education scenarios involves students making use of pre-specified criteria to assess their peers and assign marks or grades, possibly providing additional written feedback. Experimental variations of peer-

assessment design include the teacher assessing the quality of students' comments on a piece of work rather than analysing marks assigned by students to that work (Davies, 2006), students assessing each other without the provision of explicit assessment criteria (Jones & Alcock, 2014), and those designed with focus on improving specific processes in peer-assessment such as actively involving students in the development of assessment criteria in order to improve their confidence and ability in applying those criteria to assessment tasks (Smith, Cooper, & Lancaster, 2002; Orsmond, Merry, & Callaghan, 2004).

Peer-Assessment as Perceived by Students and Teachers

The perspectives of participants in peer-assessment have been sought in almost all studies that have experimented with peer-assessment.

Some studies have reported overall positive perceptions of students about being involved in peer-assessment (McLaughlin & Simpson, 2004; Saito & Fujita, 2004; Wen, Tsai, & Chang, 2006; Wen & Tsai, 2006; Kwok, 2008; Wood & Kurzel, 2008; Xiao & Lucking, 2008; McGarr & Clifford, 2013).

Some students have the view that engaging in peer-assessment tasks is productive and enables students to have an objective view of how teachers assess students (Hanrahan & Isaacs, 2001). Other advantages of peer-assessment as perceived by some students include increased responsibility for others and improved learning (Papinczak, Young, & Groves, 2007).

Views have been expressed that peer-assessment is a time-intensive process as it requires students to engage in non-trivial cognitive tasks, that it is intellectually challenging and that it creates a socially uncomfortable environment (Topping, Smith, Swanson, & Elliot, 2000; Hanrahan & Isaacs, 2001; Arnold, Shue, Kritt, Ginsburg, & Stern, 2005; Praver, Rouault, & Eidswick, 2011).

In Problem-Based Learning (PBL) environments, students have expressed their concerns that the use of peer-assessment in a summative manner may undermine the PBL practice itself, especially when feedback is not incorporated (Papinczak, Young, & Groves, 2007; Sluijsmans, Moerkerke, Van Merrienboer, & Dochy, 2001).

Students also tend to be disinclined to assessing their peers by just assigning marks and think they should provide and receive detailed and constructive feedback together with marks (Sluijsmans et al., 2001; Li & Steckelberg, 2006).

After conducting a survey of 1740 students and 460 faculty involved in peer-assessment, Liu and Carless (2006) also report that issues of reliability of peers and their perceived expertise arise when using peer-assessment in a summative manner and that most students and faculty view peer-assessment of summative nature as ineffective.

A question of high significance would be if students' negative attitudes towards peer-assessment would subside as they became more involved in peer-assessment tasks. This has been shown to be the case in one study (Sluijsmans, Brand-Gruwel, van Merriënboer, & Bastiaens, 2003), where students' levels of test anxiety decreased and their negative views of peer-assessment diminished as they progressed through three peer-assessment-based courses administered over a duration of seven months.

Psychological and Social Factors in Peer-Assessment

Gender effects are the least studied factors in peer-assessment in higher education (Falchikov & Goldfinch, 2000; Falchikov, 2003; Topping, 2010). When considering whether a student is biased by the gender of the peer they are assessing, one can safely exclude peer-assessment practices that exercise anonymity as identity information is withheld from peer-assessors. It,

however, would be interesting to find out whether there are any assessment patterns regarding gender.

The most affected peer-assessment scenarios in terms of gender are those in which the assessed work comes in the form of oral presentations. A study of 41 undergraduate students (20 females, 21 males) involved in peer-assessment of oral presentations found gender influences on the assessment process (Langan, Wheater, Shaw, Haines, Cullen, Boyle, & Preziosi, 2005). The study found that male assessors tended to rate male presenters very slightly higher than female presenters while female assessors did not show any variation in the way they assessed presenters of either gender, a finding that has been corroborated by a similar study (Langan, Shuker, Cullen, Penney, Preziosi, & Wheater, 2008).

Another study of 160 students involved in peer and self-assessment tasks (N=40 for peer assessment, 20 females, 20 males) found that female students found peer-assessment a stressing task (Pope, 2005).

Validity and Reliability of Peer-Assessment

Studies measuring the validity and reliability of peer-assessment are common in the literature. Validity of peer-assessment is measured in terms of *agreement* between scores assigned by the teacher and those assigned by students. Often referred to as reliability, inter-rater reliability measures the *closeness* of ratings by peers assessing the same piece of work or the *closeness* of the scores assigned by two or more teachers.

Fourteen studies examining the validity and reliability of peer-assessment that were published since the in-depth review by Falchikov and Goldfinch (2000) are reviewed.

In addition to the attributes adopted from the table presented by Falchikov and Goldfinch (2000, p. 8-17), two additional attributes, contribution towards final grade and anonymity, are reported.

Of the 15 studies, 8 reported correlation coefficients. Of these, 2 reported multiple correlation coefficients, which were calculated per criterion and hence were considered independent. Averages have been used to allow single comparisons with other studies.

Of the remaining 7, 4 reported standard deviations and mean, which were used to calculate effect sizes (*d*). Two of these studies had reported several standard deviation and mean values calculated for each criterion used. In those cases, the reported effect sizes are averages of the effect size calculated for each criterion.

One study (Lindblom-ylänne, Pihlajamäki, & Kotkas, 2006) did not report any statistics. Another (Cho, Schunn, & Wilson, 2006) reported correlation coefficients using bar charts, for which only approximate values could be obtained.

Two studies violated the definition of peer-assessment. The study by Ryan, Marshall, Porter and Jia (2007) exhibited a number of flaws, the most serious of which was rooted in the fact that peers were asked to assess class participation. By definition, peer-assessment involves the assessment of a piece of work produced by a student. The study in question does not involve such tasks and asking students to rate class participation is tantamount to asking them to rate students based on effort. Moreover, the study fails to report important characteristics of the experiment such as the number of students involved in the assessment task.

The study by De Grez, Valcke, and Roozen (2012) uses students from an advanced year class, who do not participate in creating the products being assessed, oral presentations, and are not rated by students. Although interesting, such assessment does not qualify as peer-assessment. Moreover, although there were 209 submissions in total, only 29 submissions were evaluated by students. The report fails to explain whether the comparison between student and teacher grades was done on those 29 presentations. Global comparison might

have used the mean and SD values reported for all 209 presentations assessed by teachers and the 29 presentations assessed by students as well. The decision to remove one teacher's evaluations from the data in order to improve agreement scores casts more doubt on the validity of the experiment for which an effect size of 1.246 has been calculated using the reported mean and SD values.

Design quality of the study by Lindblom-ylänne, Pihlajamäki, and Kotkas (2006) was deemed low because it utilised a significantly small sample size (N=15) and based its conclusions on mere comparison of mean ratings, reported using a single chart. Students were required to assess several individual dimensions instead of being instructed to use pre-specified criteria to provide global assessment. The effects of these attributes of the study could not be examined, however, because the study reported no statistics at all.

All other studies (N=12) were evaluated as having high design quality despite the fact that most of them failed to report several of the attributes discussed at the beginning of this section.

Contrasted with the number of studies included in the meta-analysis of Falchikov and Goldfinch (2000) (N>50), the number of correlation and effect size values reported here is too modest to make any strong conclusions (N=8) or to perform extensive meta-analysis. However, statistics consistent with those reported in the work of Falchikov and Goldfinch will be reported whenever possible and significant.

The correlation coefficients reported in the studies under consideration ranged from 0.396 to 0.991. As in the study by Falchikov and Goldfinch, correlation coefficients are first transformed into z-scores and the z-score of each study is weighted by the number of comparisons between teacher and student assigned marks for that study (n-comp) – 3 before the average z-score for the studies is computed and converted back into a correlation coefficient to yield the mean correlation coefficient for the studies. Statistical justifications for this conversion and for applying weights are provided by Shadish and Haddock (1994).

The mean correlation coefficient for the 8 studies calculated in this manner was $r = 0.80$. Although the number of studies considered is significantly low, this value indicates strong overall correlation between teacher and student assigned marks, and corroborates the findings of Falchikov and Goldfinch (2000).

In most of the studies discussed here, several of the design quality criteria identified by Falchikov and Goldfinch (2000) were either not met or their application was not reported. Common design pitfalls included requiring students to assess several individual dimensions instead of asking them to provide global ratings according to clearly specified criteria.

In the case of Patri (2002), unconventional control and experimental group designs such as mixing students from beginner, intermediate, and advanced level courses in the same peer-assessment experiment and allowing students to conduct group discussions regarding the work being assessed could have produced effects that are not controlled for.

In the study by Cheng and Warren (2005) three teachers assessed one class each, marking the works of 17 students on average. The significant differences in markings among the three teachers involved in the assessment suggested agreement scores among them should have been reported. These differences might have led to the variations in the effect sizes of the three classes – 0.479, -0.012, and -1.688. The higher number of students per assessment task coupled with assessment of individual dimensions may have led to fewer agreements.

The study by Cho, Schunn, and Wilson (2006) reported results using bar charts from which only ranges of correlations could be identified. The lack of reporting of exact values meant that the results of the study could not be included in the computation of the average correlation.

The study by Xiao and Lucking (2008) afforded students in the control group to remain anonymous while those in the experimental group could be identified as they were

required to disclose their identity information in the process of providing mandatory feedback. Because the study was not designed to explore the impact of anonymity, the decision to make the experiment anonymous should have been reflected across both control and treatment groups.

Other study characteristics that were either not reported or were only implied by the studies include agreements between group-assigned scores and teacher assigned scores where appropriate, population characteristics such as age and gender, contribution of peer-assessment tasks towards the final grade, anonymity, and the level of course.

Effect size (*d*) was either reported or calculated for five studies (Cheng & Warren, 2005; Ozogul & Sullivan, 2009; Matsuno, 2009; Bouzidi & Jaillet, 2009; De Grez, Valcke, & Roozen, 2012) and ranged from -0.407 to 1.246. Negative effect sizes indicate that peers were stricter than teachers and positive values indicate vice versa. The weighted average of the effect sizes, calculated using the number of comparisons as weights, was $d = 0.27$, a significant value as smaller effect sizes in peer-assessment imply more agreement between student and teacher scores (Falchikov & Goldfinch, 2000).

Due to the small number of studies (N=8 for correlation coefficient, N=5 for effect size) and missing information in some of the studies such as the number of students involved in a single assessment task and course level, it was not possible to build a descriptive linear regression model that would explain the effects of the variables under study. Nonetheless, important observations can be made by examining the data presented in table 1.

The disciplines in which the studies were conducted ranged from education, business, law, and medical education to computer science and engineering, with nearly half of the studies conducted in business and teacher education programmes. Most of the work that was assessed by peers was in the form of written assignments and oral presentations.

The study by Bouzidi and Jaillet (2009) explicitly described its goal as reducing the teacher's workload. Consequently, it sought to examine whether peer-assigned marks were tantamount to teacher assigned marks. While such an approach might be construed as contributing very little to student learning due to its strong emphasis on improving the summative value of peer-assessment, it is worth noting that, in disciplines such as computer science and mathematics, summative peer-assessment may be fairly utilised to reduce the teacher's workload as it is very likely to produce high levels of validity and reliability. In such disciplines, questions usually assess mathematical or logical reasoning, students are often required to perform calculations and develop algorithms in order to solve technical problems, and only a few and very specific criteria are used in the assessment of answers. The high correlation values reported in the study by Bouzidi and Jaillet, which involved students enrolled in a computer architecture course, may serve as evidence of this observation.

It is surprising to find that, despite recommendations based on influential reviews regarding score agreement studies, most researchers still opted to apply a single statistical method to report measurements. The statistics reported in the studies reviewed here included correlation coefficients, one-way and multiple ANOVA, Cronbach's alpha, t-tests, intra-class correlation, mean, and standard deviation. Reporting multiple statistics would allow straightforward comparison of studies and encompass the various interpretations of validity and reliability in the process.

Discussion

While research in peer-assessment has been conducted in both academic and professional settings, the studies reviewed here were all conducted in higher education settings. Regardless, the variables of interest to each study and the settings in which it was conducted have led to a multitude of peer-assessment design strategies, most of which are commendable

and provide insight into the intricacies of the practice. Studies by Cho et al. (2006), Ozogual and Sullivan (2009), Smith et al. (2002) , and Xiao and Lucking (2008) are exemplary for involving a large cohort of students while Sahin (2008) highlights the advantage of involving students in the specification and development of assessment criteria.

Although a fair proportion of the studies preserve anonymity of students, the challenges of doing so are highlighted as the number of students grows. Moreover, processes such as oral presentations and interviews can hardly be anonymised. Yet, the advantages of anonymity, whenever it can be applied, shall not be underestimated as it has the potential to minimise undesired behaviour such as favouritism or bias.

Several issues of concern in peer-assessment have been found to reverberate across the wide array of studies investigated.

Lack of common standards for peer-assessment practices stands out among these issues as it has made the evaluation and comparison of peer-assessment practices and instruments difficult if not impossible. Researchers have yet to agree on exact interpretations of validity and reliability of scores, which statistics to use to measure and report agreement scores, and most importantly, on how peer-assessment experiments should be set up and conducted. Most studies mix experiments and attempt to measure several variables using one-off experiments.

Another issue of concern is that many peer-assessment practices have failed to take advantage of advances in related disciplines. Although a few studies have pointed out how peer-assessment can be incorporated into comprehensive learning environments such as Problem-Based Learning (PBL) and Collaborative Learning, the fact remains that the vast majority of peer-assessment activities are standalone practices built on top of traditional classrooms.

Advances in computer science disciplines are being applied in almost all social systems and scientific disciplines to help solve problems that were deemed intractable or very challenging until recently. Unfortunately, peer-assessment has yet to take advantage of such advances as the use of computers in the discipline has not gone beyond implementing web-based peer-assessment tools. In an upcoming study, the author intends to provide a review of the problems in peer-assessment that practitioners deem challenging and to demonstrate how similar problems in computer science have been solved or are currently being addressed.

The majority of peer-assessment practices are conducted in a one-off or non-iterative manner. The validity of this approach is put in doubt when the goal of the peer-assessment task is to measure how the practice improves long-term learning. Such learning outcome cannot possibly be measured over one or two semesters. In fact, its effective measurement involves putting in place programmes that implement peer-assessment throughout the duration of the educational programme itself.

The requirements for introducing such programmes in higher education institutions are however restrictive as they involve redesigning well-established curricula, additional investment, taking considerable risks both on parts of the institution and students, and may require making modifications to existing policies. This is probably the most prohibitive reason that has limited practitioners to implementing peer-assessment for shorter durations and in usually small class sizes.

Despite this restriction, a large number of peer-assessment studies have been conducted over the past fifteen years. The disconcerting fact, however, is that most of these studies are disconnected and only a few truly build upon previous findings. It appears that most studies have insignificant variations in the variables being studied and usually reach similar conclusions that neither strengthen nor contradict the findings of previous studies. Given the restrictive nature of the problem, the most productive path for researchers to follow would be to conduct incremental research or research that replicates previous findings if solid

results are to be established. This observation is probably best revealed by comparing how peer-assessment score agreement studies that have been conducted during the past fifteen years are strikingly similar, both in their design and findings, to those conducted in the previous century.

Other peer-assessment factors that have been identified by scholars as needing further investigation but have received relatively small attention include the impact of gender, race, and similar factors on the process, how anonymity plays a role in lessening or eliminating unintended effects of these factors, how to address possible educational dishonesty such as plagiarism and collusive behaviour, and impact of formative peer-assessment on the performance of students in tasks of summative nature such as final exams.

Formative peer-assessment could help students monitor their own progress and identify their strengths and weaknesses. For the teacher, formative peer-assessment may also serve the purpose of identifying and monitoring students that may need additional supervision. The potential role of formative peer-assessment as a tool of early intervention is, however, not investigated in many of the studies. It is understood that this role can hardly be studied in one-off experiments and its investigation essentially requires redesigning these experiments as iterative processes. Nonetheless, researchers interested in exploring the applicability of formative peer-assessment are encouraged to consider exploring this potential role as one of the motives for designing future experiments with iterative and replicable processes.

Manual peer-assessment is common in many of these studies. The opinions of teachers involved in many of these studies reflect that manual peer-assessment is just as burdensome as traditional assessment while most students identify the unfair increase in workload as a potential deterrent of practicing peer-assessment. Automation has already proved successful in reducing the workload of both students and teachers as well as in eliminating other unintended problems brought about by manual peer-assessment such as bias and favouritism. Researchers might argue that their specific peer-assessment design is difficult to automate but it should be noted that all the peer-assessment designs applied in the studies discussed here can be automated, although to varying degrees.

Recommendations

The author recommends a number of possible additions to future research in peer-assessment. One is to explore the applicability of educational games to peer-assessment practices. The application of educational games in traditional classrooms has been under investigation for over fifty years. Whether they actually improve student learning is still open to debate and findings vary across fields. Some early studies found simulation games showed little or no superiority to traditional instruction in the social sciences (Cohen & Bradley, 1978; Fraas, 1980; Szafran & Mandolini, 1980; Klein & Freitag, 1991) whereas positive results were reported in the fields of Math, Physics, and biology (DeVries & Slavin, 1976; White, 1984; Spraggins & Rowsey, 1986). Yet, it should be noted that most early studies focused on the applicability of educational games at elementary and high school levels. For a thorough review of studies that introduced educational games to traditional classrooms and investigated their effectiveness, the reader is invited to see Randel, Morris, Wetzel, and Whitehill (1992) and Wu, Chiou, Kao, Hu, and Huang (2012).

Although inconclusive results and questionability of studies regarding the applicability of educational games to traditional classrooms may be suggestive of similar outcomes in the introduction of such games to the realm of peer-assessment, researchers are strongly encouraged to consider the degree to which advances in computer science and related technologies have had an immense role in overcoming many challenges in both

academia and industry when contemplating the potential of educational games to enhance peer-assessment.

Peer-assessment is a scenario in which two or more students are involved in completing tasks that require fairly equivalent levels of participation for the entire process to be effective. Eliciting participation when students are not willing to actively participate in learning activities, a case not specific to peer-assessment, usually involves providing incentives of either collaborative or competitive nature to enhance the learning process.

Recent studies have investigated the effectiveness of using competitive and collaborative games to improve learning outcomes and increase student involvement. The most notable of these is the study by Burguillo (2010), which utilises game theory to build competitive tournaments in which groups of students from a computer science course compete at the end of the course. In this tournament, groups of students compete in the Prisoner's Dilemma game (Axelrod & Hamilton, 1981) to earn extra points that will count towards their final scores for the course. Based on positive and consistent student survey results over five editions of the same course in which the tournaments were conducted, Burgillo suggests that competitive games provide strong motivation for students and increase their performance. Other recent studies seeking to augment the learning process with competitive games have also reported positive results (Lawrence, 2004; Pareto, Haake, Lindström, Sjödén, & Gulz, 2012; Muñoz-Merino, Molina, Muñoz-Organero, & Kloos, 2012; Hwang, Wu, & Chen, 2012; Mustika, Sari, Kao, & Heh, 2014).

The values of interaction and collaboration among students as part of the learning process have been emphasised in recent studies as well. Many of these studies utilise computer software to enhance the collaboration process. An earlier study conducted among 127 MBA students that used a group decision support system to enhance collaboration reported that the process led to higher levels of skill development and learning as perceived by students as well as better performance at end-of-course exams (Alavi, 1994). More recent technology-based collaborative learning studies involve Internet-based learning environments (Sun & Shen, 2014; Rojas, Kapralos, & Dubrowski, 2014; Michailidis & Tsiatsos, 2014).

Automation of peer-assessment tasks could therefore allow researchers to efficiently incorporate healthy competition and collaboration into the practice and conduct further research on the impact of these variables.

Automation of peer-assessment practices could also greatly enhance the efficiency of the processes involved. One of the possible reasons for not designing manual peer-assessment tasks as iterative processes is that they would be time consuming and ultimately impractical. For instance, random distribution of peer-assessment tasks, coupled with multiple rounds of feedback would become impossible as the number of students involved grows. Automated peer-assessment can be designed to be virtually free of any delay that is introduced as a result of manual distribution of assignments and communication among students. Indeed, automation streamlines the processes to allow efficient and iterative communication among peers in the provision of feedback and revision and resubmission of the assessed work (Gehringer, 2001; Sitthiworachart & Joy, 2003; Li, Steckelberg & Srinivasan, 2008).

Other advantages that come with automation of peer-assessment tasks include moving towards a ubiquitous learning environment where peer-assessment is not confined to the classroom (Jones & Jo, 2004; Sun & Shen, 2014) and reduced teacher workload (Bouzidi, & Jaillet, 2009).

Advanced opportunities brought about by automation of peer-assessment tasks are automated detection of academic dishonesty such as plagiarism, application of social network analysis in large classes, automated essay scoring, automatic calibration of peer-assigned scores (Hamer, Ma, & Kwong, 2005; Giovannella & Scaccia, 2014), and utilising student-

generated data to build models that predict student performance (Ashenafi, Riccardi, & Ronchetti, 2015).

Students, especially in the initial stages of peer-assessment, are often critical of their peers' ability in assessing their work. Another recommendation is research on whether this criticism has foundation or arises from bias. A possible scenario is where a teacher plays the role of a student and assesses 'peers' in an anonymous peer-assessment experiment where students are not notified of the teacher's involvement. Changes in opinions of students, or otherwise, after the conclusion of the experiment shall provide enough information to accept or reject the null hypothesis that students are not unreasonably critical of their peers' ability to assess their work.

Many positive findings have come from peer-assessment studies conducted over the past fifteen years. For instance, most of the studies discussed have shown that although student have doubts and initially tend to resist being involved in peer-assessment, such resistance subsides over time. Most of these studies also support the findings reflected in reviews of studies conducted before the turn of the century.

However, peer-assessment is at a stage where practitioners and educators need to establish design quality and measurement standards for it to emerge from the forest of solitary case studies and small-scale short-term experiments to become the revolutionary educational assessment practice it has long promised to be. The establishment of such standards guarantees proper evaluation and comparison of practices and promotes novel and incremental research by specifying clear milestones and roadmaps.

It is also an opportune time for scholars in education and computer science as well as for other practitioners of peer-assessment to realise that peer-assessment is now an interdisciplinary practice. Some of the challenges that prevent large-scale implementation and detailed study of peer-assessment practices already have their counterparts in computer science either solved or under rigorous study. Therefore, interfaculty collaborations will be just as important as the establishment of standards in allowing researchers to focus on the most important factors in order to bring to fruition peer-assessment practices of the 21st century.

References

Alavi, M. (1994). Computer-Mediated Collaborative Learning: An Empirical Evaluation. *MIS Quarterly, 18*(2), 159–174. http://doi.org/10.2307/249763

Althauser, R., & Darnall, K. (2001). Enhancing Critical Reading and Writing through Peer Reviews: An Exploration of Assisted Performance. *Teaching Sociology*, 23-35. http://doi.org/10.2307/1318780

Arnold, L., Shue, C. K., Kritt, B., Ginsburg, S., & Stern, D. T. (2005). Medical students' views on peer assessment of professionalism. *Journal of General Internal Medicine*, *20*(9), 819–824. http://doi.org/10.1111/j.1525-1497.2005.0162.x

Ashenafi, M. M., Riccardi, G., & Ronchetti, M. (in press). Predicting Students' Final Exam Scores from their Course Activities. In *Frontiers in Education, 2015*. IEEE.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science (New York, N.Y.), 211*(4489), 1390–6. http://doi.org/10.1126/science.7466396

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education, 27*(5), 427-441. http://doi.org/10.1080/0260293022000009302

Bangert-Drowns, R. L., Wells-Parker, E., & Chevillard, I. (1997). Assessing the methodological quality of research in narrative reviews and meta-analyses. In *Assessing Methodological Quality* (pp. 405–429). http://psycnet.apa.org/books/10222/012

Bloxham*, S., & West, A. (2004). Understanding the rules of the game: marking peer assessment as a medium for developing students' conceptions of assessment. *Assessment & Evaluation in Higher Education, 29*(6), 721-733. http://doi.org/10.1080/0260293042000227254

Bouzidi, L., & Jaillet, A. (2009). Can online peer assessment be trusted? *Educational Technology and Society, 12*(4), 257–268. http://www.ifets.info/journals/12_4/22.pdf

Burguillo, J. C. (2010). Using game theory and competition-based learning to stimulate student motivation and performance. *Computers & Education, 55*(2), 566-575. http://dx.doi.org/10.1016/j.compedu.2010.02.018

Campbell, K. S., Mothersbaugh, D. L., Brammer, C., & Taylor, T. (2001). Peer versus Self-Assessment of Oral Business Presentation Performance. *Business Communication Quarterly, 64*(3), 23-40. http://doi.org/10.1177/108056990106400303

Chen, Y., & Tsai, C. (2009). An educational research course facilitated by online peer assessment. *Innovations in Education and Teaching International, 46*(1), 105-117. http://doi.org/10.1080/14703290802646297

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing, 22*(1), 93-121. http://doi.org/10.1191/0265532205lt298oa

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction*, *20*(4), 328–338. http://doi.org/10.1016/j.learninstruc.2009.08.006

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*(4), 891. http://doi.org/10.1037/0022-0663.98.4.891

Cohen, R. B., & Bradley, R. H. (1978). Simulation games, learning, and retention. *The Elementary School Journal, 78*(4), 247-253. http://www.jstor.org/stable/1001260

Davies, P. (2006). Peer assessment: judging the quality of students' work by comments rather than marks. *Innovations in Education and Teaching International, 43*(1), 69-82. http://doi.org/10.1080/14703290500467566

De Grez, L., Valcke, M., & Roozen, I. (2012). How effective are self- and peer assessment of oral presentation skills compared with teachers' assessments? *Active Learning in Higher Education, 13*(2), 129-142. http://doi.org/10.1177/1469787412441284

DeVries, D. L., & Slavin, R. E. (1976). Teams-games-tournament: a final report on the research. (Report No. 217). Baltimore, MD: John Hopkins University, Center for the Study of Sical Organization of Schools.

Falchikov, N. (2003). Involving Students in Assessment. *Psychology Learning & Teaching, 3*(2), 102-108. http://doi.org/10.2304/plat.2003.3.2.102

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. http://doi.org/10.3102/00346543070003287

Fraas, J. W. (1980). The use of seven simulation games in a college economics course. *The Journal of Experimental Education, 48*(4), 264-280.

Gehringer, E. F. (2001). Electronic peer review and peer grading in computer-science courses. *ACM SIGCSE Bulletin, 33*(1), 139-143.

Gielen, S., Dochy, F., Onghena, P., Struyven, K., & Smeets, S. (2011). Goals of peer assessment and their associated quality concepts. *Studies in Higher Education, 36*(6), 719-735. http://doi.org/10.1080/03075071003759037

Giovannella, C., & Scaccia, F. (2014, July). Technology-Enhanced" Trusted" Participatory Grading. In *Advanced Learning Technologies (ICALT), 2003 IEEE 14th International Conference on* (pp. 347-349). IEEE.

Hamer, J., Ma, K. T., & Kwong, H. H. (2005, January). A method of automatic grade calibration in peer assessment. In *Proceedings of the 7th Australasian conference on Computing education* (Vol. 42, pp. 67-72).

Hanrahan, S. J., & Isaacs, G. (2001). Assessing Self- and Peer-assessment: The students' views. *Higher education research and development, 20*(1), 53-70. http://doi.org/10.1080/07294360123776

Harlen, W., & James, M. (1997). Assessment and Learning: differences and relationships between formative and summative assessment. *Assessment in Education, 4*(3), 365-379. http://doi.org/10.1080/0969594970040304

Hu, G. (2005). Using peer review with Chinese ESL student writers. *Language Teaching Research, 9*(3), 321-342. http://doi.org/10.1191/1362168805lr169oa

Hwang, G. J., Wu, P. H., & Chen, C. C. (2012). An online game approach for improving students' learning performance in web-based problem-solving activities. *Computers & Education, 59*(4), 1246-1256.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774-1787.

Jones, V., & Jo, J. H. (2004, December). Ubiquitous learning environment: An adaptive teaching system using ubiquitous technology. In *Beyond the comfort zone: Proceedings of the 21st ASCILITE Conference* (Vol. 468, p. 474).

Klein, J. D., & Freitag, E. (1991). Effects of using an instructional game on motivation and performance. *The Journal of Educational Research, 84*(5), 303-308.

Kollar, I., & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive perspective. *Learning and Instruction*, *20*(4), 344–348. http://doi.org/10.1016/j.learninstruc.2009.08.005

Kwok, L. (2008). Students' perception of peer evaluation and teachers' role in seminar discussions. *Electronic journal of foreign language teaching, 5*(1), 84-97.

Langan, A. M., Shuker, D. M., Cullen, W. R., Penney, D., Preziosi, R. F., & Wheater, C. P. (2008). Relationships between student characteristics and self-, peer and tutor evaluations of oral presentations. *Assessment & Evaluation in Higher Education, 33*(2), 179-190. http://doi.org/10.1080/02602930701292498

Langan*, A. M., Wheater, C. P., Shaw, E. M., Haines, B. J., Cullen, W. R., Boyle, J. C., ... & Preziosi, R. F. (2005). Peer assessment of oral presentations: effects of student gender, university affiliation and participation in the development of assessment criteria. *Assessment & Evaluation in Higher Education, 30*(1), 21-34.

Lawrence, R. (2004). Teaching data structures using competitive games. *Education, IEEE Transactions on, 47*(4), 459-466.

Li, L., & Steckelberg, A. L. (2006). Perceptions of web-mediated peer assessment. *Academic Exchange Quarterly, 10*(2), 265.

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. http://doi.org/10.1111/j.1467-8535.2009.00968.x

Li, L., Steckelberg, A. L., & Srinivasan, S. (2008). Utilizing peer interactions to promote learning through a web-based peer assessment system. *Canadian Journal of Learning and Technology/La revue canadienne de l'apprentissage et de la technologie, 34*(2).

Lin, S. S. J., Liu, E. Z. F., & Yuan, S. M. (2001). Web-based peer assessment: Feedback for students with various thinking-styles. *Journal of Computer Assisted Learning*, *17*(4), 420–432. http://doi.org/10.1046/j.0266-4909.2001.00198.x

Lindblom-ylänne, S., Pihlajamäki, H., & Kotkas, T. (2006). Self-, peer- and teacher-assessment of student essays. *Active Learning in Higher Education, 7*(1), 51-62. http://doi.org/10.1177/1469787406061148

Liu, N. F., & Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher education, 11*(3), 279-290.

Luxton-Reilly, A. (2009). A systematic review of tools that support peer assessment. *Computer Science Education, 19*(4), 209-232. http://doi.org/10.1080/08993400903384844

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 075-100. http://doi.org/10.1177/0265532208097337

McGarr, O., & Clifford, A. M. (2013). 'Just enough to make you take it seriously': exploring students' attitudes towards peer assessment. *Higher education, 65*(6), 677-693.

McLaughlin, P., & Simpson, N. (2004). Peer assessment in first year university: How the students feel. *Studies in Educational Evaluation, 30*(2), 135-149.

Michailidis, N., & Tsiatsos, T. (2014, July). Supporting Students by Using Interaction Analysis Tools in Educational Group Blogging: A Case Study of the GIANT Tool. In *Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on* (pp. 291-293). IEEE.

Miller, P. J. (2003). The Effect of Scoring Criteria Specificity on Peer and Self-assessment. *Assessment & Evaluation in Higher Education, 28*(4), 383-394. http://doi.org/10.1080/0260293032000066218

Min, H. T. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing*, *15*(2), 118–141. http://doi.org/10.1016/j.jslw.2006.01.003

Morgan, C., & O'reilly, M. (1999). *Assessing open and distance learners.* Psychology Press.

Muñoz-Merino, P. J., Molina, M. F., Muñoz-Organero, M., & Kloos, C. D. (2012). An adaptive and innovative question-driven competition-based intelligent tutoring system for learning. *Expert Systems with Applications, 39*(8), 6932-6948.

Mustika, M., Sari, M. L., Kao, C. T., & Heh, J. S. (2014, July). Digital BINGO Game as a Dynamic Assessment in a Reading Instruction for Learning Indonesian as a Foreign Language: A System Architecture. In *Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on* (pp. 219-221). IEEE.

Myers, S. (2008). Formative and Summative Assessments. *Research Starters Education (Online Edition)*

Orsmond*, P., Merry, S., & Callaghan, A. (2004). Implementation of a formative assessment model incorporating peer and self-assessment. *Innovations in Education and Teaching International, 41*(3), 273-290. http://doi.org/10.1080/14703290410001733294

Otoshi, J., & Heffernan, N. (2007). An analysis of peer assessment in EFL college oral presentation classrooms. *Language Teacher-Kyoto-JALT, 31*(11), 3.

Ozogul, G., & Sullivan, H. (2009). Student performance and attitudes under formative evaluation by teacher, self and peer evaluators. *Educational Technology Research and Development*, *57*(3), 393–410. http://doi.org/10.1007/s11423-007-9052-7

Papinczak, T., Young, L., & Groves, M. (2007). Peer assessment in problem-based learning: A qualitative study. *Advances in Health Sciences Education*, *12*(2), 169–186. http://doi.org/10.1007/s10459-005-5046-6

Pareto, L., Haake, M., Lindström, P., Sjödén, B., & Gulz, A. (2012). A teachable-agent-based game affording collaboration and competition: evaluating math comprehension and motivation. *Educational Technology Research and Development, 60*(5), 723-751.

Patri, M. (2002). The influence of peer feedback on self-and peer-assessment of oral skills. *Language Testing, 19*(2), 109-131.

Pope*, N. K. L. (2005). The impact of stress in self- and peer assessment. *Assessment & Evaluation in Higher Education, 30*(1), 51-63. http://doi.org/10.1080/0260293042003243896

Praver, M., Rouault, G., & Eidswick, J. (2011). Attitudes and affect toward peer evaluation in EFL reading circles. *The Reading Matrix*, *11*(2), 89–101.

Randel, J. M., Morris, B. A., Wetzel, C. D., & Whitehill, B. V. (1992). The effectiveness of games for educational purposes: A review of recent research. *Simulation & gaming, 23*(3), 261-276.

Rojas, D., Kapralos, B., & Dubrowski, A. (2014, July). Gamification for Internet Based Learning in Health Professions Education. In *Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on* (pp. 281-282). IEEE.

Rudy, D. W., Fejfar, M. C., Griffith, C. H., & Wilson, J. F. (2001). Self- and peer assessment in a first-year communication and interviewing course. *Evaluation & the Health Professions*, *24*(4), 436–445. http://doi.org/10.1177/016327870102400405

Ryan, G. J., Marshall, L. L., Porter, K., & Jia, H. (2007). Peer, professor and self-evaluation of class participation. *Active Learning in Higher Education, 8*(1), 49-61. http://doi.org/10.1177/1469787407074049

Sahin, S. (2008). An Application of Peer Assessment in Higher Education. *The Turkish Online Journal of Educational Technology, 7*(2).

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing, 25*(4), 553-581. http://doi.org/10.1177/0265532208094276

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research, 8*(1), 31-54. http://doi.org/10.1191/1362168804lr133oa

Shadish, W. R., & Haddock, C. (1994). Combining estimates of effect size. *The handbook of research synthesis* (pp. 261–281).

Sitthiworachart, J., & Joy, M. (2003, July). Web-based peer assessment in learning computer programming. In *Advanced Learning Technologies (ICALT), 2003 IEEE 14th International Conference on* (p. 180). IEEE.

Sluijsmans, D. M. A., Brand-Gruwel, S., van Merriënboer, J. J. G., & Bastiaens, T. J. (2003). The training of peer assessment skills to promote the development of reflection skills in teacher education. *Studies in Educational Evaluation*, *29*(1), 23–42. http://doi.org/10.1016/S0191-491X(03)90003-4

Sluijsmans, D. M. A., Brand-gruwel, S., van Merriënboer, J. J. G., & Martens, R. L. (2004). Training teachers in peer-assessment skills: effects on performance and perceptions [La formation des enseignants par l'évaluation des compétences entre pairs]. *Innovations in Education and Teaching International*, *41*(1), 59–78.

Sluijsmans, D. M. A., Moerkerke, G., van Merrienboer, J. J. G., & Dochy, F. J. R. C. (2001). Peer Assessment in Problem Based Learning. *Studies in educational evaluation, 27*(2), 153-173. http://doi.org/10.1016/S0191-491X(01)00019-0

Sluijsmans, D., & Prins, F. (2006). A conceptual framework for integrating peer assessment in teacher education. *Studies in Educational Evaluation*, *32*(1), 6–22. http://doi.org/10.1016/j.stueduc.2006.01.005

Smith, H., Cooper, A., & Lancaster, L. (2002). Improving the Quality of Undergraduate Peer Assessment: A Case for Student and Staff Development. *Innovations in Education and Teaching International, 39*(1), 71-81. http://doi.org/10.1080/13558000110102904

Spraggins, C. C., & Rowsey, R. E. (1986). The effect of simulation games and worksheets on learning of varying ability groups in a high school biology classroom. *Journal of Research in Science Teaching, 23*(3), 219-229.

Strijbos, J. W., & Sluijsmans, D. (2010). Unravelling peer assessment: Methodological, functional, and conceptual developments. *Learning and Instruction*, *20*(4), 265–269. http://doi.org/10.1016/j.learninstruc.2009.08.002

Strijbos, J. W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction*, *20*(4), 291–303. http://doi.org/10.1016/j.learninstruc.2009.08.008

Sun, G., & Shen, J. (2014, July). Collaborative learning through TaaS: a mobile system for courses over the cloud. In *Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on* (pp. 278-280). IEEE.

Szafran, R. F., & Mandolini, A. F. (1980). Test performance and concept recognition: The effect of a simulation game on two types of cognitive knowledge. Simulation & Games.

Tillema, H., Leenknecht, M., & Segers, M. (2011). Assessing assessment quality: Criteria for quality assurance in design of (peer) assessment for learning - A review of research studies. *Studies in Educational Evaluation*, *37*(1), 25–34. http://doi.org/10.1016/j.stueduc.2011.03.004

Topping, K. J. (2010). Methodological quandaries in studying process and outcomes in peer assessment. *Learning and Instruction*, *20*(4), 339–343. http://doi.org/10.1016/j.learninstruc.2009.08.003

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative Peer Assessment of Academic Writing Between Postgraduate Students. *Assessment & Evaluation in Higher Education, 25*(2), 149-169. http://doi.org/10.1080/713611428

Topping, K.J. (1998). Peer Assessment Between Students in Colleges and Universities. *Review of educational Research, 68*(3), 249-276. http://doi.org/10.3102/00346543068003249

Tsai, C. C., Lin, S. S., & Yuan, S. M. (2002). Developing science activities through a networked peer assessment system. *Computers & Education, 38*(1), 241-252. http://doi.org/10.1016/S0360-1315(01)00069-0

Van den Berg, I., Admiraal, W., & Pilot, A. (2006a). Designing student peer assessment in higher education: analysis of written and oral peer feedback. *Teaching in Higher Education, 11*(2), 135-147. http://doi.org/10.1080/13562510500527685

Van den Berg, I., Admiraal, W., & Pilot, A. (2006b). Peer assessment in university teaching: evaluating seven course designs. *Assessment & Evaluation in Higher Education, 31*(1), 19-36. http://doi.org/10.1080/02602930500262346

Van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educational Research Review, 4*(1), 41-54. http://doi.org/10.1016/j.edurev.2008.11.002

Van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, *20*(4), 270–279. http://doi.org/10.1016/j.learninstruc.2009.08.004

Wen, M. L., & Tsai, C. C. (2006). University students' perceptions of and attitudes toward (online) peer assessment. *Higher Education, 51*(1), 27-44. http://doi.org/10.1007/s10734-004-6375-8

Wen, M. L., Tsai, C. C., & Chang, C. Y. (2006). Attitudes towards peer assessment: a comparison of the perspectives of pre-service and in-service teachers. *Innovations in Education and Teaching International, 43*(1), 83-92. http://doi.org/10.1080/14703290500467640

White, B. Y. (1984). Designing computer games to help physics students understand Newton's laws of motion. *Cognition and instruction, 1*(1), 69-108.

Wood, D., & Kurzel, F. (2008). Engaging students in reflective practice through a process of formative peer review and peer assessment. *ATN Assessment, 1*(1).

Wu, W. H., Chiou, W. B., Kao, H. Y., Hu, C. H. A., & Huang, S. H. (2012). Re-exploring game-assisted learning research: The perspective of learning theoretical bases. *Computers & Education, 59*(4), 1153-1161.

Xiao, Y., & Lucking, R. (2008). The impact of two types of peer assessment on students' performance and satisfaction within a Wiki environment. *Internet and Higher Education*, *11*(3-4), 186–193. http://doi.org/10.1016/j.iheduc.2008.06.005

Table 1

Teacher-peer score agreement studies and their attributes

Study	Population characteristics	Subject area and course name	What is assessed and level	Instrument and criteria	Design quality	Statistics reported	Value of comparison metrics	Number involved per assessment	Contribution to final grade	Anonymity preserved?
Lin et al. (2001)	n-part=n-comp=58 18 female, 40 male	Computer science Operating systems course	Written assignment Introductory?	6 specific criteria + Holistic feedback, 10-point Likert scale	H	Correlation coefficient	*r = 0.396* for rating after feedback	2 teachers 6 students	Not stated	Yes
Campbell et al. (2001)	n-part=n-comp=66 21-47 years old, 61% female, 85% Caucasian, 0-30 years of work experience	Business communication course	Oral team presentations Intermediate	Holistic rating and analytical rating criteria six 5-point scales, three 5-point holistic rating scales	H	Correlation coefficient	*r = 0.45*, average of 5 criteria	1 teacher ? students	0%	Not stated
Rudy Fejfar, Griffith, & Wilson (2001)	n-part=97 n-eff=n-comp=82	Medical education Interviewing course	Interviewing performance Introductory?	Three criteria Three 15-point Likert scale items	H	Pearson Correlation coefficient	*r=0.50* df = 86, p = .0001 for composite score ratings	1 teacher ≈ 8 students	0%	Yes

Study	Population characteristics	Subject area and course name	What is assessed and level	Instrument and criteria	Design quality	Statistics reported	Value of comparison metrics	Number involved per assessment	Contribution to final grade	Anonymity preserved?
Patri (2002)	n-part=56 n-eff=n-comp=54 18-21 years old Control group (n=29) Experimental group (n=25)	Multiple departments English foundation programme (n=41) Practice speaking for communication (n=13)	Oral presentation Introductory) Speaking Practice (level not stated)	Teacher-specified criteria Fourteen 5-point Likert scale questions	H	Correlation coefficient	r=0.49 for control group *r=0.85* for experimental group	1 teacher 3-4 students	0%	No
Cheng & Warren (2005)	n-part=n-comp=51 49 male, 2 female	Electrical engineering – English for academic purposes course	Seminar, oral presentation, written report Introductory?	Teacher-specified criteria Twelve 5-point Likert scale questions	H	Mean and SD Paired t-test	The average of average effect sizes over 12 criteria for the three classes is reported: *d = -0.407*	1 teacher 12-17 students	20%	Not stated
Lindblom-ylänne et al. (2006)	n-part=n-comp=15, 1 male	Law, course on the history of law	Critical essay Level not stated	Teacher specified criteria Seven 4-point scale ratings	L	Graphical report of ratings	No statistical analysis performed, simple comparison of means	1 teacher 1 student	100% but students not told of the decision until the end of the assessment tasks	Yes

Study	Population characteristics	Subject area and course name	What is assessed and level	Instrument and criteria	Design quality	Statistics reported	Value of comparison metrics	Number involved per assessment	Contribution to final grade	Anonymity preserved?
Cho et al. (2006)	n-part=708, n-comp=272 61% female	16 courses	Written assignment Mixed course levels	Teacher-specified criteria Three evaluation dimensions with 7-point scale ratings	H	Pearson correlation Root Mean Squared Error Intra-class correlation Standard deviation	Instructor and students' views of validity and reliability reported using charts	1 teacher 4-6 students	Typically about 40%	Yes
Otoshi & Heffernan (2007)	n-part=n-comp=67 50 male	Economics and business administration	presentation level not stated	Teacher-specified criteria, six dimensions measured using 5-point Likert scale	H	Cronbach's alpha, Mean, SD, Correlation coefficient	Cronbach's alpha .82 for class 1 and .79 for class 2 averages reported class 1: r = 0.663 class 2: r = 0.609	1 teacher 31 or 36 students	Not stated	Not stated
Ryan et al. (2007)	n-eff=96 24.5 years old on average 89 females 63 Caucasian, 14 African American, 19 Other	4 courses	Class participation Advanced?	A single 4-point scale criterion (class participation)	L	Bias and precision, Pearson's correlation	Overall bias: 0.48 Overall precision 36%	1 teacher ? students	20-25%	Not stated

Study	Population characteristics	Subject area and course name	What is assessed and level	Instrument and criteria	Design quality	Statistics reported	Value of comparison metrics	Number involved per assessment	Contribution to final grade	Anonymity preserved?
Sahin (2008)	n-part=n-comp=48	Education Specific teaching methods I course	Team-project Advanced?	Students involved in the specification of assessment criteria Thirty 4-point criteria	H	Mean, Mode, Median SD, Skewness, Kurtosis, range, Pearson correlation	*r = 0.991*, p<0.01	1 teacher ? students	Not stated	Yes
Xiao & Lucking (2008)	n-part=232, n-comp=230 77% Caucasian, 79.5% female 25.04 years old on average	Teacher education Introductory course on social and cultural foundations of American education	A 1000-word article Introductory?	Teacher-specified criteria Four 5-point Likert scale items	H	Intra-class correlation, Pearson's correlation	ICC for first round: r=.62, p<.05 ICC for second round: r=.75, p<.001 Agreement: *r(230)=.829*, p<.001	1 teacher 3-4 students or 20 students	5%	No
Bouzidi & Jaillet (2009)	Group 1: n-part-n-comp=68, 36 male Group 2: n-part=n-comp=94, 42 male	Computer science, two editions of a computer architecture course	Written exams, 2nd and 3rd year students Intermediate?	Marking instructions and scales provided by the teacher, items marked from 0.5 to 3, with 0.5 levels of increment	H	Pearson's correlation, T-Test, Effect size r	r1=0.88, r2=0.89 Effect size r1=0.06 Effect size r2=0.02	1 teacher 4 students	Yes but percentage not reported	Yes

Study	Population characteristics	Subject area and course name	What is assessed and level	Instrument and criteria	Design quality	Statistics reported	Value of comparison metrics	Number involved per assessment	Contribution to final grade	Anonymity preserved?
Ozogul & Sullivan (2009)	n-part=n-comp=133	Teacher education programme, computer in education course	Lesson plan, scored posttest Introductory	10-item lesson plan evaluation rubric provided by researcher	H	Mean, SD, Effect size (*d*) calculated	Effect size calculated from mean and SD of scores by peers and researcher *d = 0.320*	1 teacher 1 student	55%	Yes
Matsuno (2009)	n-part=n-comp=97 19 to 21 years old	Two university writing classes	Essay level not stated	Teacher-specified criteria Sixteen 6-point Likert scale essay evaluation criteria	H	Effect sizes (*d*) calculated	Average effect size reported *d = 0.380*	2 teachers 5 students	10%	Yes
De Grez et al. (2012)	n-part=n-comp=57 21 female, 18 years old on average	Business administration	oral presentation Introductory	Teacher-specified criteria Nine 5-point Likert scale items	L	Intra-class correlation, mean and SD, effect size calculated	*d = 1.246*	5 teachers 6 students	Not stated	No

n-part = number of participants, n-comp = number of comparisons, n-eff = actual number of participants after some students were removed from the class