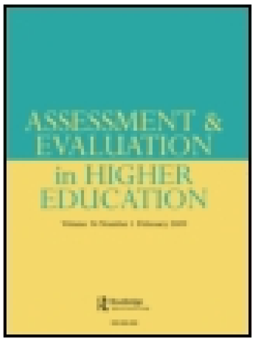

Routledge
Taylor & Francis Group

Assessment & Evaluation in Higher Education

ISSN: 0260-2938 (Print) 1469-297X (Online) Journal homepage: http://www.tandfonline.com/loi/caeh20

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

Michael Mogessie Ashenafi

To cite this article: Michael Mogessie Ashenafi (2015): Peer-assessment in higher education – twenty-first century practices, challenges and the way forward, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2015.1100711

To link to this article: http://dx.doi.org/10.1080/02602938.2015.1100711

Published online: 19 Oct 2015.

Assessment & Evaluation in Higher Education, 2015
http://dx.doi.org/10.1080/02602938.2015.1100711

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

Michael Mogessie Ashenafi*

Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

Peer assessment in higher education has been studied for decades. Despite the substantial amount of research carried out, peer assessment has yet to make significant advances. This review identifies themes of recent research 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 and James 1997; Morgan and O'Reilly 1999; Myers 2014). 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 and James 1997; Morgan and O'Reilly 1999).

Formative assessment is rather student-centred – it is intended to provide 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.

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 assessments. In these environments, students are heavily involved in evaluating their works as well as those of their peers.

*Email: michael.mogessie@unitn.it

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' (250).

A significant amount of research in the area of peer assessment has been conducted since the turn of the century. The purpose of this study is to provide a comprehensive review of this research 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 twenty-first 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 it 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.

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:

- 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.

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 assessment of professional practice rather than academic products and processes, were excluded from the analysis. Because this study is intended to extend existing reviews (Topping 1998; Falchikov and 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 64 articles.

Literature review

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 issues in peer assessment that have been identified by scholars and practitioners.

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, should allow replication and provide students with clear instructions regarding the processes involved. The importance of student involvement in all stages is also highlighted by Tillema, Leenknecht, and Segers (2011), while the importance of involving students in the specification of assessment criteria has also been stressed by other studies (Bloxham and West 2004; Sluijsmans et al. 2004).

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 skills, peer assessment skills and students' attitudes towards peer assessment are four variable categories Van Zundert, Sluijsmans, and van 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, 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 assessment is individual or group-based both for assessor and assessed, whether assessment is in-class or out-of-class, and whether there is a reward for participation.

In a later study, the authors experiment with these variables to determine their impact on oral and written feedback (Van den Berg, Admiraal, and 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 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 for task interdependence, which refers to the responsible involvement of students in peer assessment tasks, the interpersonal variables identified are specific to group-based peer assessment tasks.

Quality of peer assessment

Reviews show that, although peer assessment quality is often discussed, there have been few studies that evaluate the quality of peer assessment methods, and that quality standards have yet to be formally defined as practitioners 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* – 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 assessment tasks being 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 et al. (2011), who demonstrate that the quality criteria that should be met in peer assessment are determined by the goal of the 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 practices would meet a single set of quality measurement criteria.

Case studies, action research and peer assessment instruments

The studies in this category investigate specific settings 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 showed that elaborate and specific feedback from peers, although found adequate, was perceived as having a negative impact by the students receiving such feedback (Strijbos, Narciss, and 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 that specific feedback is more helpful than holistic feedback in improving students' performance.

It is also claimed that those students who provide their peers with high-quality feedback tend to incorporate feedback from their peers effectively, raising their final grades in the process (Althauser and Darnall 2001; Tsai, Lin, and Yuan 2002). While another study could not confirm the existence of this relationship, it found a significant relationship between the quality of feedback a student provided and the quality of their own final project (Li, Liu, and Steckelberg 2010). How the nature of 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 that 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 in 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 (Hu 2005; Min 2006; Sluijsmans and Prins 2006; Saito 2008). According to Chen and Tsai (2009), however, the improvement in quality becomes less significant in subsequent sessions.

Peer assessment design strategies

In a class of 12 students enrolled in a two-year postgraduate course, Topping et al. (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 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 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 settings.

This study demonstrates what many peer assessment practices share, and yet how dissimilar they are, in the sense that the findings of one study support or contradict those of another. A relatively small number of students, one-off experiments, incomparable results and two 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 prespecified 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 et al. preserves anonymity and uses well-designed criteria, but its findings are 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, this study, which utilised 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 manual, and 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 they 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 can afford to enjoy the advantages that come with peer assessment without the negative impacts discussed, because such tools offer anonymity and can easily automate assignment distribution, discussion and submission of feedback and grades. Automated assessment can also help with calibrating grades assigned by multiple peers (Hamer, Ma, and Kwong 2005).

The most common implementation of peer assessment in higher education scenarios involves students making use of prespecified criteria to assess their peers and assign marks or grades, possibly providing additional written feedback. Experimental variations of 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 and Alcock 2014), and those focusing 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 (Smith, Cooper, and Lancaster 2002; Orsmond, Merry, and Callaghan 2004).

Peer assessment as perceived by students and teachers

The perspectives of participants in peer assessment have been sought in almost all studies. Some studies have reported overall positive perceptions of students (McLaughlin and Simpson 2004; Saito and Fujita 2004; Wen and Tsai 2006; Wen, Tsai, and Chang 2006; Kwok 2008; Wood and Kurzel 2008; Xiao and Lucking 2008; McGarr and Clifford 2013). Some students have the view that engaging in peer assessment tasks is productive and enables them to have a clearer view of how teachers assess students (Hanrahan and Isaacs 2001). Other advantages of peer assessment as perceived by students include increased responsibility for others and improved learning (Papinczak, Young, and Groves 2007).

It has also 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 et al. 2000; Hanrahan and Isaacs 2001; Arnold et al. 2005; Praver, Rouault, and Eidswick 2011). In problem-based learning environments, students have expressed their concerns that the use of peer assessment in a summative manner may undermine learning, especially when feedback is not incorporated (Sluijsmans et al. 2001; Papinczak, Young, and Groves 2007). Students also tend to be disinclined to assess their peers by just assigning marks, and think they should provide and receive detailed and constructive feedback (Sluijsmans et al. 2001; Li and 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 this as ineffective.

Do students' negative attitudes towards peer assessment subside as they became more involved in peer-assessment tasks? This has been shown to be the case in one study (Sluijsmans et al. 2003), where students' levels of test anxiety decreased and their negative views 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 and 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 practices that exercise anonymity.

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 oral presentations found gender influences on the assessment process (Langan et al. 2005). 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 et al. 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 it a stressful task (Pope 2005).

Validity and reliability of peer assessment

Studies measuring validity and reliability are common. Validity 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) were reviewed. In addition to the attributes adopted from the table presented by Falchikov and Goldfinch (8–17), two additional attributes, contribution towards final grade and anonymity, are reported.

Of the 15 studies, eight reported correlation coefficients. Of these, two 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 seven, four reported standard deviations and mean, which were used to calculate effect sizes (d). Two of these studies reported 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, and Kotkas 2006) did not report any statistics. Another (Cho, Schunn, and 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 et al. (2007) exhibited a number of flaws, the most serious of which was 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.

The study by De Grez, Valcke, and Roozen (2012) uses students from an advanced year class, who did not participate in creating the products being assessed, oral presentations, which 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 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.

Design quality of the study by Lindblom-ylänne, Pihlajamäki, and Kotkas (2006) was deemed low, because it had a 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 prespecified 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, though most of them failed to report several of the attributes discussed at the beginning of this section. By comparison 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, 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 eight studies calculated in this manner was $r = 0.80$. Although the number of studies considered is 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 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 marking 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) allowed 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 and Warren 2005; Bouzidi and Jaillet 2009; Matsuno 2009; Ozogul and Sullivan 2009; De Grez, Valcke, and 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 imply more agreement between student and teacher scores (Falchikov and 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 analysis of variance, Cronbach's alpha, *t*-tests, intraclass 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.

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, Liu, and Yuan (2001)	n-part = n-comp = 58	Computer science	Written assignment	6 specific criteria + holistic feedback, 10-point Likert scale	H	Correlation coefficient	$r = 0.396$ for rating after feedback	2 teachers	Not stated	Yes
	18 female, 40 male	Operating systems course	Introductory?					6 students		
Campbell et al. (2001)	n-part = n-comp = 66	Business communication course	Oral team presentations	Holistic rating and analytical rating criteria six fivr-point scales, three fivr-point holistic rating scales	H	Correlation coefficient	$r = 0.45$, average of 5 criteria	1 teacher	0%	Not stated
	21–47 years old, 61% female, 85% Caucasian, 0–30 years of work experience		Intermediate					? students		
Rudy et al. (2001)	n-part = 97	Medical education interviewing course	Interviewing performance	Three criteria	H	Pearson	$r = 0.50$ df = 86, $p = 0.0001$ for composite score ratings	1 teacher ≈ 8 students	0%	Yes
	n-eff = n-comp = 82		Introductory?	Three 15-point Likert scale items		Correlation coefficient				

(*Continued*)

Table 1. (*Continued*).

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?
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	Multiple departments	Oral presentation Introductory)	Teacher-specified criteria	H	Correlation coefficient	$r = 0.49$ for control group	1 teacher	0%	No
	n-eff = n-comp = 54	English foundation programme ($n = 41$)	Speaking practice (level not stated)	Fourteen five-point Likert scale questions			$r = 0.85$ for experimental group	3–4 students		
	18–21 years old	Practice speaking for communication ($n = 13$)								
	Control group ($n = 29$) Experimental group ($n = 25$)									
Cheng and Warren (2005)	n-part = n-comp = 51	Electrical engineering – English for academic purposes course	Seminar, oral presentation, written report	Teacher-specified criteria	H	Mean and SD	The average of average effect sizes over 12 criteria for the three classes is reported: $d = -0.407$	1 teacher	20%	Not stated
	49 male, 2 female		Introductory?			Paired *t*-test		12–17 students		

				Twelve five-point Likert scale questions						
Lindblom-ylänne, Pihlajamäki, and Kotkas (2006)	n-part = n-comp = 15, 1 male	Law, course on the history of law	Critical essay	Teacher-specified criteria	L	Graphical report of ratings	No statistical analysis performed, simple comparison of means	1 teacher	100% but students not told of the decision until the end of the assessment tasks	Yes
			Level not stated	Seven four-point scale ratings				1 student		
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, Schunn, and Wilson (2006)	n-part = 708, n-comp = 272, 61% female	16 courses	Written assignment	Teacher-specified criteria	H	Pearson correlation	Instructor and students' views of validity and reliability reported using charts	1 teacher	Typically about 40%	Yes
			Mixed course levels	Three evaluation dimensions with seven-point scale ratings		Root Mean Squared Error Intra-class correlation		4–6 students		

(*Continued*)

Table 1. (*Continued*).

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?
						Standard deviation				
Otoshi and Heffernan (2007)	n-part = n-comp = 67	Economics and business administration	Presentation level not stated	Teacher-specified criteria, six dimensions measured using five-point Likert scale	H	Cronbach's alpha, Mean, SD, Correlation coefficient	Cronbach's alpha 0.82 for class 1 and 0.79 for class 2	1 teacher	Not stated	Not stated
	50 male						Averages reported class 1: $r = 0.663$ class 2: $r = 0.609$	31 or 36 students		
Ryan et al. (2007)	n-eff = 96	4 courses	Class participation	A single four-point scale criterion (class participation)	L	Bias and precision, Pearson's correlation	Overall bias: 0.48	1 teacher	20–25%	Not stated
	24.5 years old on average 89 females 63 Caucasian, 14 African American, 19 other		Advanced?				Overall precision 36%	? students		

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	Team-project	Students involved in the specification of assessment criteria	H	Mean, Mode, Median SD, Skewness, Kurtosis, range, Pearson correlation	$r = 0.991$, $p < 0.01$	1 teacher	Not stated	Yes
		Specific teaching methods I course	Advanced?	Thirty four-point criteria				? students		
Xiao and Lucking (2008)	n-part = 232, n-comp = 230	Teacher education	A 1000-word article	Teacher-specified criteria	H	Intra-class correlation, Pearson's correlation	ICC for first round: $r = 0.62$, $p < 0.05$	1 teacher	5%	No
	77% Caucasian, 79.5% female	Introductory course on social and cultural foundations of American education	Introductory?	Four five-point Likert scale items			ICC for second round: $r = 0.75$, $p < 0.001$	3–4 students or 20 students		
	25.04 years old on average						Agreement: r (230) = 0.829, $p < 0.001$			

(*Continued*)

Table 1. (*Continued*).

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?
Bouzidi and Jaillet (2009)	Group 1: n-part-n-comp = 68, 36 male	Computer science, two editions of a computer architecture course	Written exams, 2nd and 3rd year students	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	$r1 = 0.88$, $r2 = 0.89$	1 teacher	Yes but percentage not reported	Yes
	Group 2: n-part = n-comp = 94, 42 male		Intermediate?				Effect size $r1 = 0.06$	4 students		
							Effect size $r2 = 0.02$			
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 and Sullivan (2009)	n-part = n-comp = 133	Teacher education programme, computer in education course	Lesson plan, scored post-test	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	1 teacher	55%	Yes
			Introductory				$d = 0.320$	1 student		
Matsuno (2009)	n-part = n-comp = 97	Two university writing classes	Essay level not stated	Teacher-specified criteria	H	Effect sizes (*d*) calculated	Average effect size reported	2 teachers	10%	Yes
	19–21 years old						$d = 0.380$	5 students		

				Sixteen six-point Likert scale essay evaluation criteria						
De Grez, Valcke, and Roozen (2012)	n-part = n-comp = 57	Business administration	Oral presentation	Teacher-specified criteria	L	Intra-class correlation, mean and SD, effect size calculated	*d* = 1.246	5 teachers	Not stated	No
	21 female, 18 years old on average		Introductory	Nine five-point Likert scale items				6 students		

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

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. The variables of interest to each study and the settings in which they were conducted have led to a multitude of design strategies, most of which are commendable and provide insight into the intricacies of the practice. Studies by Cho, Schunn, and Wilson (2006), Ozogul and Sullivan (2009), Smith, Cooper, and Lancaster (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, should not be underestimated as it has the potential to minimise undesired behaviour such as favouritism or bias.

Several issues of concern have been found to reverberate across the wide array of studies investigated. Lack of common standards stands out among these issues, as it has made the evaluation and comparison of 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.

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 and collaborative learning, the vast majority of peer assessment activities are standalone practices in conventional 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 has not gone beyond implementing web-based 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 task is to measure how the practice improves long-term learning. Such learning outcomes cannot possibly be measured over one or two semesters. 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 small class sizes.

Despite this restriction, a large number of studies have been conducted over the past fifteen years. Most of these studies are disconnected and only a few truly build upon previous findings. 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 15 years are strikingly similar, both in their design and findings, to those conducted in the previous century.

Other 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 the impact of formative peer assessment on the performance of students in tasks of summative nature such as final examinations.

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 who 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 designing future experiments with iterative and replicable processes.

Manual peer assessment is common in many studies. The opinions of teachers involved indicate that manual assessment is just as burdensome as conventional assessment, while most students identify the unfair increase in workload as a potential deterrent to 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 design is difficult to automate, but it should be noted that all the 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 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 conventional instruction in the social sciences (Cohen and Bradley 1978; Fraas 1980; Szafran and Mandolini 1980;

Klein and Freitag 1991), whereas positive results were reported in the fields of mathematics, physics and biology (DeVries and Slavin 1976; White 1984; Spraggins and 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, the reader is referred to Randel et al. (1992) and Wu et al. (2012).

Although the results 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 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 and 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; Hwang, Wu, and Chen 2012; Muñoz-Merino et al. 2012; Pareto et al. 2012; Mustika et al. 2014).

The values of interaction and collaboration among students as part of the learning process have also been emphasised in recent studies. 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 examinations (Alavi 1994). More recent technology-based collaborative learning studies involve Internet-based learning environments (Michailidis and Tsiatsos 2014; Rojas, Kapralos, and Dubrowski 2014; Sun and Shen 2014).

Automation of peer assessment tasks could allow researchers to efficiently incorporate healthy competition and collaboration into the practice, and conduct further research on the impact of these variables. Automation 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 and Joy 2003; Li, Steckelberg, and 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 and Jo 2004; Sun and Shen 2014), and reduced teacher workload (Bouzidi and Jaillet 2009). Advanced opportunities brought about by automation include 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, and Kwong 2005; Giovannella and Scaccia 2014), and utilising student-generated data to build models that predict student performance (Ashenafi, Riccardi, and Ronchetti in press).

Students, especially in the initial stages, 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 experiment where students are not notified of the teacher's involvement. Changes in opinions of students, or otherwise, after the conclusion of the experiment should 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, 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 it 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 twenty-first century.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Michael Mogessie Ashenafi is a doctoral candidate with the department of computer science at the University of Trento. His research focuses on the application of advanced computer algorithms to student-generated data to automate peer-assessment practices. He has recently noted that there is lack of collaboration between scholars in education and their counterparts in computer science in addressing the challenges of peer- assessment. He has hence carried out this review with the intention of demonstrating that little progress in peer-assessment research has been made since the turn of the century and that an interdisciplinary approach is needed in order to move forward.

References

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

Althauser, R., and K. Darnall. 2001. “Enhancing Critical Reading and Writing through Peer Reviews: An Exploration of Assisted Performance.” *Teaching Sociology* 29 (1): 23–35. doi:10.2307/1318780.

Arnold, L., C. K. Shue, B. Kritt, S. Ginsburg, and D. T. Stern. 2005. “Medical Students’ Views on Peer Assessment of Professionalism.” *Journal of General Internal Medicine* 20 (9): 819–824. doi:10.1111/j.1525-1497.2005.0162.x.

Ashenafi, M. M., G. Riccardi, and M. Ronchetti. in press. “Predicting Students’ Final Exam Scores from their Course Activities.” In *Frontiers in Education*. El Paso, TX: IEEE. 20–25 Oct 2015.

Axelrod, R., and W. D. Hamilton. 1981. “The Evolution of Cooperation.” *Science* 211 (4489): 1390–1396. doi:10.1126/science.7466396.

Ballantyne, R., K. Hughes, and A. Mylonas. 2002. “Developing Procedures for Implementing Peer Assessment in Large Classes using an Action Research Process.” *Assessment & Evaluation in Higher Education* 27 (5): 427–441. doi:10.1080/0260293022000009302.

Bangert-Drowns, R. L., E. Wells-Parker, and I. Chevillard. 1997. “Assessing the Methodological Quality of Research in Narrative Reviews and Meta-analyses.” In *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research*, 405–429. Washington, DC: American Psychological Association, xxxii, 458 pp. doi:10.1037/10222-012.

Bloxham, S., and A. West. 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. doi:10.1080/0260293042000227254.

Bouzidi, L., and A. Jaillet. 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. doi:10.1016/j.compedu.2010.02.018.

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

Chen, Y., and C. Tsai. 2009. “An Educational Research Course Facilitated by Online Peer Assessment.” *Innovations in Education and Teaching International* 46 (1): 105–117. doi:10.1080/14703290802646297.

Cheng, W., and M. Warren. 2005. “Peer Assessment of Language Proficiency.” *Language Testing* 22 (1): 93–121. doi:10.1191/0265532205lt298oa.

Cho, K., and C. MacArthur. 2010. “Student Revision with Peer and Expert Reviewing.” *Learning and Instruction* 20 (4): 328–338. doi:10.1016/j.learninstruc.2009.08.006.

Cho, K., C. D. Schunn, and R. W. Wilson. 2006. “Validity and Reliability of Scaffolded Peer Assessment of Writing from Instructor and Student Perspectives.” *Journal of Educational Psychology* 98 (4): 891–901. doi:10.1037/0022-0663.98.4.891.

Cohen, R. B., and R. H. Bradley. 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. doi:10.1080/14703290500467566.

DeVries, D. L., and 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 Social Organization of Schools.

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

Falchikov, N., and J. Goldfinch. 2000. “Student Peer Assessment in Higher Education: A Meta-analysis Comparing Peer and Teacher Marks.” *Review of Educational Research* 70 (3): 287–322. doi: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., F. Dochy, P. Onghena, K. Struyven, and S. Smeets. 2011. “Goals of Peer Assessment and their Associated Quality Concepts.” *Studies in Higher Education* 36 (6): 719–735. doi:10.1080/03075071003759037.

Giovannella, C., and F. Scaccia. 2014, July. “Technology-enhanced ‘Trusted’ Participatory Grading.” In *2014 IEEE 14th International Conference on Advanced Learning Technologies (ICALT)*, Athens, 347–349. IEEE.

De Grez, L., M. Valcke, and I. Roozen. 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. doi:10.1177/1469787412441284.

Hamer, J., K. T. Ma, and H. H. Kwong. 2005, January. “A Method of Automatic Grade Calibration in Peer Assessment.” In *Proceedings of the 7th Australasian Conference on Computing Education*, Newcastle, Vol. 42, 67–72.

Hanrahan, S. J., and G. Isaacs. 2001. “Assessing Self- and Peer-assessment: The Students’ Views.” *Higher Education Research & Development* 20 (1): 53–70. doi:10.1080/07294360123776.

Harlen, W., and M. James. 1997. “Assessment and Learning: Differences and Relationships between Formative and Summative Assessment.” *Assessment in Education: Principles, Policy & Practice* 4 (3): 365–379. doi:10.1080/0969594970040304.

Hu, G. 2005. “Using Peer Review with Chinese ESL Student Writers.” *Language Teaching Research* 9 (3): 321–342. doi:10.1191/1362168805lr169oa.

Hwang, G. J., P. H. Wu, and C. C. Chen. 2012. “An Online Game Approach for Improving Students’ Learning Performance in Web-based Problem-solving Activities.” *Computers & Education* 59 (4): 1246–1256.

Jones, I., and L. Alcock. 2014. “Peer Assessment Without Assessment Criteria.” *Studies in Higher Education* 39 (10): 1774–1787.

Jones, V., and J. H. Jo. 2004, December. “Ubiquitous Learning Environment: An Adaptive Teaching System using Ubiquitous Technology.” In *Beyond the Comfort Zone: Proceedings of the 21st ASCILITE Conference*, Perth, Vol. 468, 474.

Klein, J. D., and E. Freitag. 1991. “Effects of using an Instructional Game on Motivation and Performance.” *The Journal of Educational Research* 84 (5): 303–308.

Kollar, I., and F. Fischer. 2010. “Peer Assessment as Collaborative Learning: A Cognitive Perspective.” *Learning and Instruction* 20 (4): 344–348. doi: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., D. M. Shuker, W. R. Cullen, D. Penney, R. F. Preziosi, and C. P. Wheater. 2008. “Relationships between Student Characteristics and Self-, Peer and Tutor Evaluations of Oral Presentations.” *Assessment & Evaluation in Higher Education* 33 (2): 179–190. doi:10.1080/02602930701292498.

Langan, A. M., C. P. Wheater, E. M. Shaw, B. J. Haines, W. R. Cullen, J. C. Boyle, and R. F. Preziosi. 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.” *IEEE Transactions on Education* 47 (4): 459–466.

Li, L., X. Liu, and A. L. Steckelberg. 2010. “Assessor or Assessee: How Student Learning Improves by Giving and Receiving Peer Feedback.” *British Journal of Educational Technology* 41 (3): 525–536. doi:10.1111/j.1467-8535.2009.00968.x.

Li, L., and A. L. Steckelberg. 2006. “Perceptions of Web-mediated Peer Assessment.” *Academic Exchange Quarterly* 10 (2): 265.

Li, L., A. L. Steckelberg, and S. Srinivasan. 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): 1–10.

Lin, S. S. J., E. Z. F. Liu, and S. M. Yuan. 2001. “Web-based Peer Assessment: Feedback for Students with Various Thinking-styles.” *Journal of Computer Assisted Learning* 17 (4): 420–432. doi:10.1046/j.0266-4909.2001.00198.x.

Lindblom-ylänne, S., H. Pihlajamäki, and T. Kotkas. 2006. “Self-, Peer- and Teacher-assessment of Student Essays.” *Active Learning in Higher Education* 7 (1): 51–62. doi:10.1177/1469787406061148.

Liu, N. F., and D. Carless. 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. doi:10.1080/08993400903384844.

Matsuno, S. 2009. “Self-, Peer-, and Teacher-assessments in Japanese University EFL Writing Classrooms.” *Language Testing* 26 (1): 075–100. doi:10.1177/0265532208097337.

McGarr, O., and A. M. Clifford. 2013. “‘Just Enough to Make You Take It Seriously’: Exploring Students’ Attitudes towards Peer Assessment.” *Higher Education* 65 (6): 677–693.

McLaughlin, P., and N. Simpson. 2004. “Peer Assessment in First Year University: How the Students Feel.” *Studies in Educational Evaluation* 30 (2): 135–149.

Michailidis, N., and T. Tsiatsos. 2014, July. “Supporting Students by using Interaction Analysis Tools in Educational Group Blogging: A Case Study of the GIANT Tool.” In *2014 IEEE 14th International Conference on Advanced Learning Technologies (ICALT)*, Athens, 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. doi: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. doi:10.1016/j.jslw.2006.01.003.

Morgan, C., and M. O’Reilly. 1999. *Assessing Open and Distance Learners*. London: Psychology Press.

Muñoz-Merino, P. J., M. F. Molina, M. Muñoz-Organero, and C. D. Kloos. 2012. “An Adaptive and Innovative Question-driven Competition-based Intelligent Tutoring System for Learning.” *Expert Systems with Applications* 39 (8): 6932–6948.

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

Myers, S. 2014. “Formative & Summative Assessments.” *Research Starters Education* (Online Edition). Retrieved from http://connection.ebscohost.com/c/essays/27577770/formative-summative-assessments

Orsmond, P., S. Merry, and A. Callaghan. 2004. “Implementation of a Formative Assessment Model Incorporating Peer and Self-assessment.” *Innovations in Education and Teaching International* 41 (3): 273–290. doi:10.1080/14703290410001733294.

Otoshi, J., and N. Heffernan. 2007. “An Analysis of Peer Assessment in EFL College Oral Presentation Classrooms.” *The Language Teacher-Kyoto-JALT* 31 (11): 3–8.

Ozogul, G., and H. Sullivan. 2009. “Student Performance and Attitudes under Formative Evaluation by Teacher, Self and Peer Evaluators.” *Educational Technology Research and Development* 57 (3): 393–410. doi:10.1007/s11423-007-9052-7.

Papinczak, T., L. Young, and M. Groves. 2007. "Peer Assessment in Problem-based Learning: A Qualitative Study." *Advances in Health Sciences Education* 12 (2): 169–186. doi:10.1007/s10459-005-5046-6.

Pareto, L., M. Haake, P. Lindström, B. Sjödén, and A. Gulz. 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. doi:10.1080/0260293042003243896.

Praver, M., G. Rouault, and J. Eidswick. 2011. "Attitudes and Affect toward Peer Evaluation in EFL Reading Circles." *The Reading Matrix* 11 (2): 89–101.

Randel, J. M., B. A. Morris, C. D. Wetzel, and B. V. Whitehill. 1992. "The Effectiveness of Games for Educational Purposes: A Review of Recent Research." *Simulation & Gaming* 23 (3): 261–276.

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

Rudy, D. W., M. C. Fejfar, C. H. Griffith, and J. F. Wilson. 2001. "Self- and Peer Assessment in a First-year Communication and Interviewing Course." *Evaluation & the Health Professions* 24 (4): 436–445. doi:10.1177/016327870102400405.

Ryan, G. J., L. L., Marshall, K. Porter, and H. Jia. 2007. "Peer, Professor and Self-evaluation of Class Participation." *Active Learning in Higher Education* 8 (1): 49–61. doi:10.1177/1469787407074049.

Sahin, S. 2008. "An Application of Peer Assessment in Higher Education." *The Turkish Online Journal of Educational Technology* 7 (2): 5–10.

Saito, H. 2008. "EFL Classroom Peer Assessment: Training Effects on Rating and Commenting." *Language Testing* 25 (4): 553–581. doi:10.1177/0265532208094276.

Saito, H., and T. Fujita (2004). "Characteristics and User Acceptance of Peer Rating in EFL Writing Classrooms." *Language Teaching Research* 8 (1): 31–54. doi:10.1191/1362168804lr133oa.

Shadish, W. R., and C. Haddock. 1994. "Combining Estimates of Effect Size." In *The Handbook of Research Synthesis*, edited by H. Cooper, L.V. Hedges, 261–281. New York: Russell Sage Foundation.

Sitthiworachart, J., and M. Joy. 2003, July. "Web-based Peer Assessment in Learning Computer Programming." In *2003 IEEE 3rd International Conference on Advanced Learning Technologies (ICALT)*, Athens, 180–184. IEEE.

Sluijsmans, D. M. A., S. Brand-Gruwel, J. J. G. van Merriënboer, and T. J. Bastiaens. 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. doi:10.1016/S0191-491X(03)90003-4.

Sluijsmans, D. M. A., S. Brand-Gruwel, J. J. G. van Merriënboer, and R. L. Martens. 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., G. Moerkerke, J. J. G. van Merrienboer, and F. J. R. C. Dochy. 2001. "Peer Assessment in Problem Based Learning." *Studies in Educational Evaluation* 27 (2): 153–173. doi:10.1016/S0191-491X(01)00019-0.

Sluijsmans, D., and F. Prins. 2006. "A Conceptual Framework for Integrating Peer Assessment in Teacher Education." *Studies in Educational Evaluation* 32 (1): 6–22. doi:10.1016/j.stueduc.2006.01.005.

Smith, H., A. Cooper, and L. Lancaster. 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. doi:10.1080/13558000110102904.

Spraggins, C. C., and R. E. Rowsey. 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., S. Narciss, and K. Dünnebier. 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. doi:10.1016/j.learninstruc.2009.08.008.

Strijbos, J. W., and D. Sluijsmans. 2010. “Unravelling Peer Assessment: Methodological, Functional, and Conceptual Developments.” *Learning and Instruction* 20 (4): 265–269. doi:10.1016/j.learninstruc.2009.08.002.

Sun, G., and J. Shen. 2014, July. “Collaborative Learning through TaaS: A Mobile System for Courses Over the Cloud.” In *2014 IEEE 14th International Conference on Advanced Learning Technologies (ICALT)*, Athens, 278–280. IEEE.

Szafran, R. F., and A. F. Mandolini. 1980. “Test Performance and Concept Recognition: The Effect of a Simulation Game on Two Types of Cognitive Knowledge.” *Simulation & Games* 11 (3): 326–335.

Tillema, H., M. Leenknecht, and M. Segers. 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. doi:10.1016/j.stueduc.2011.03.004.

Topping, K. J. 1998. “Peer Assessment between Students in Colleges and Universities.” *Review of Educational Research* 68 (3): 249–276. doi:10.3102/00346543068003249.

Topping, K. J. (2010). “Methodological Quandaries in Studying Process and Outcomes in Peer Assessment.” *Learning and Instruction* 20 (4): 339–343. doi:10.1016/j.learninstruc.2009.08.003.

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

Tsai, C. C., S. S. Lin, and S. M. Yuan. 2002. “Developing Science Activities through a Networked Peer Assessment System.” *Computers & Education* 38 (1): 241–252. doi:10.1016/S0360-1315(01)00069-0.

Van den Berg, I., W. Admiraal, and A. Pilot. 2006a. “Designing Student Peer Assessment in Higher Education: Analysis of Written and Oral Peer Feedback.” *Teaching in Higher Education* 11 (2): 135–147. doi:10.1080/13562510500527685.

Van den Berg, I., W. Admiraal, and A. Pilot. 2006b. “Peer Assessment in University Teaching: Evaluating Seven Course Designs.” *Assessment & Evaluation in Higher Education* 31 (1): 19–36. doi:10.1080/02602930500262346.

Van Gennip, N. A. E., M. S. R. Segers, and H. H. Tillema. 2009. “Peer Assessment for Learning from a Social Perspective: The Influence of Interpersonal Variables and Structural Features.” *Educational Research Review* 4 (1): 41–54. doi:10.1016/j.edurev.2008.11.002.

Van Zundert, M., D. Sluijsmans, and J. van Merriënboer. 2010. “Effective Peer Assessment Processes: Research Findings and Future Directions.” *Learning and Instruction* 20 (4): 270–279. doi:10.1016/j.learninstruc.2009.08.004.

Wen, M. L., and C. C. Tsai. 2006. “University Students’ Perceptions of and Attitudes toward (Online) Peer Assessment.” *Higher Education* 51 (1): 27–44. doi:10.1007/s10734-004-6375-8.

Wen, M. L., C. C. Tsai, and C. Y. Chang. 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. doi: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., and F. Kurzel. 2008. “Engaging Students in Reflective Practice through a Process of Formative Peer Review and Peer Assessment.” *ATN Assessment Conference 2008*. Retrieved from http://www.ojs.unisa.edu.au/index.php/atna/article/download/376/252

Wu, W. H., W. B. Chiou, H. Y. Kao, C. H. A. Hu, and S. H. Huang. 2012. “Re-exploring Game-assisted Learning Research: The Perspective of Learning Theoretical Bases.” *Computers & Education* 59 (4): 1153–1161.

Xiao, Y., and R. Lucking. 2008. “The Impact of Two Types of Peer Assessment on Students’ Performance and Satisfaction within a Wiki Environment.” *The Internet and Higher Education* 11 (3–4): 186–193. doi:10.1016/j.iheduc.2008.06.005.