Self- and peer-assessment: Evidence from the Accounting and Finance discipline

Omaima A. G. Hassan, Alison Fox and Gwen Hannah

Address for correspondence:

Dr Omaima A. G. Hassan

Department of Economics & Finance,

Brunel University,

Uxbridge,

UB8 3PH

Omaima.hassan@brunel.ac.uk

Self- and peer-assessment: Evidence from the Accounting and
Finance discipline

Abstract

Self- and peer-assessment of student work is an area that is under-researched in the accounting education literature, although the subject area of study seems to influence the results obtained in prior studies. The current study contributes to the literature by examining the accuracy and construct validity of self- and peer-assessment by accounting students. It also investigates students' views about these exercises. The findings show that whilst the self- and peer-assessment of students appear to be neither accurate nor valid, the students are positive about the impact of these procedures on their learning experience. These findings indicate that, although instructors might not rely on self- and peer- assessment as measures of student performance for the purpose of summative assessment, the exercise may prove useful for formative assessment because it can promote a wide range of transferable skills.

Key words: Self- assessment; peer- assessment; accuracy; validity; students' views.

1. Introduction

Continuous technological and economic changes in the business environment have increased the demand for lifelong learners who should be reflective, self-directed, original thinkers and who can contribute to society at large (Boud, Cohen and Sampson, 1999). Both self-assessment (SA) and peer- assessment (PA) are considered to be useful tools in the development of lifelong learning, as they help promote a wide range of transferable skills (Boud and Lublin, 1983; Stefani, 1994). It is argued that SA, (the ability of learners to assess and evaluate their own work) enables students to monitor, direct and regulate their actions towards the goals of information acquisition, increased expertise and self-improvement (Lew, Alwis and Schmidt, 2010). It has also been suggested that higher education should build on students' existing abilities to assess their own work and generate their own feedback (Nicol and Macfarlane-Dick, 2006). Similarly, PA (an assessment of students' work by their peers) gives students a greater sense of ownership and empathy for the subjective judgements required during the assessment process (Ellington, 1996). These arguments suggest that both SA and PA should be actively used in educational practice (Falchicov and Goldfinch, 2000); however, in higher education this is often not the case as both formative[1] and summative[2] assessments are generally seen to be the responsibility of those who teach (Nicol and Macfarlane-Dick, 2006).

Although the existing SA and PA literature is extensive and covers many disciplines such as medicine, dentistry and foreign languages (e.g., Stefani, 1994; Orsmond, Merry and Reiling, 1996; Longhurst and Norton,1997; Fitzgerald, White and

Gruppen, 2003; Cho, Schunn and Wilson, 2006; Lindblom-ylanne, Pihlajamaki and Kotkas, 2006; Evans, Leeson and Petrie, 2007; Papinczak et al., 2007; Wen and Tsai, 2008; Matsuno, 2009; De Grez, Valcke and Berings, 2010; Lew, Alwis and Schmidt, 2010; Liang and Tsai, 2010), it has produced mixed results about the accuracy and construct validity of SA and PA. Therefore, this area of research remains a viable topic of investigation. In addition, although the subject area of study seems to have contributed to the results obtained from prior studies (Falchikov and Boud, 1989; Falchikov and Goldfinch, 2000; Dochy, Segers and Sluijsmans, 1999), less than a handful of SA and/or PA studies are relevant to accounting education, some having been conducted in the business studies discipline (e.g., Orpen[3], 1994; Freeman, 1995; Johnson and Smith[4], 1997; Larres, Ballantine and Whittington, 2003). These few studies vary in terms of scope (either self- or peer- evaluations), purpose (examine the accuracy, reliability and/or validity of the assessment) and the student task (presentations, quizzes or tests). The current study contributes to the literature by examining the accuracy and construct validity of both SA and PA in two accounting education contexts and analyses students' views about the usefulness of these exercises. It differs from existing research because it uses both quantitative and qualitative methods to assess the effectiveness of SA and PA as a development tool for lifelong learning. The findings show that both SA and PA are not accurate when compared to a teacher's assessment (TA); there are significant differences between the students' marks and a teacher's mark. However, PA is moderately valid, which means that both students and teachers are roughly ranking students in a similar order. The construct validity of SA is found to be less robust, however. These findings indicate that PA and SA may not

4

be appropriate for the purpose of summative assessment. However, the results of the student survey show that this should not prevent the implementation of SA and PA for the purpose of formative assessment, since both are considered valuable by students in their learning process.

The remainder of this paper is organised as follows. Section Two provides a review of the literature relating to the accuracy and construct validity of SA and PA; it also discusses students' perceptions of these exercises. Section Three outlines the research methods employed whilst Section Four discusses the data collection and presents an analysis of the results. Finally, Section Five provides concluding remarks.

2. Literature review

2.1 Self- and peer- assessment

Although many prior studies assess the accuracy and/or validity of SA and/or PA, these terms are often not distinguished from one another; indeed, many studies use the terms interchangeably[5] (e.g., Larres, Ballantine and Whittington, 2003; Evans, Leeson and Petrie, 2007). Therefore, the current study provides a clear definition of each term in order to avoid any such confusion. Some define accuracy as the closeness of the agreement between a value obtained by a measurement and the true value of the thing being measured (International Vocabulary of Basic and General Terms in Metrology, 2006, p.46). This definition of accuracy is often adopted in the SA and/or PA literature; for example, Lynn, Holzer and O'Neill (2006) and Regehr and Eva (2006) define the

accuracy of SA as the degree of agreement between the marks attributed by both the student and an expert about the relative rank of the student's skills. Alternatively, Blanch-Hartigan (2011) defines the accuracy of SA as the direct relationship or comparison between self-evaluation scores and an objective or expert evaluation that serves as the standard for performance assessment (for instance TA). This definition is appealing because it is not limited to the degree of agreement between a measure and a reference point but allows comparisons to be made. The current study adopts Blanch-Hartigan's (2011) definition of accuracy in the context of SA and PA using paired comparisons, that is SA and/or PA are/is said to be accurate if there are no significant differences between these assessments and TA.

Indeed, a number of studies have examined the accuracy of SA and/or PA (e.g. Fitzgerald, White and Gruppen, 2003; Larres, Ballantine and Whittington, 2003; Freeman, 1995; Longhurst and Norton, 1997; Papinczak et al., 2007; Lew, Alwis and Schmidt, 2010). For example, Fitzgerald, White and Gruppen (2003) contrast students' own estimated performance (SA) with their actual performance (TA), to examine the stability of SA accuracy over time for undergraduate medical students. They develop three measures of SA accuracy: (i) bias (arithmetic differences in actual and estimated scores); (ii) deviation (absolute differences in actual and estimated scores); and (iii) covariation (correlation of actual and estimated scores) using students' actual performance as a reference point. These authors conclude that the correlation between an actual and an estimated score is not a very useful indicator of SA accuracy because it quantifies an aspect that is distinct from those summarized in the bias and deviation measures. In one of the few studies conducted in accounting education, Larres,

Ballantine and Whittington (2003) use the Wilcoxon matched-pairs signed-rank test to examine the differences between SA and actual performance of entry-level undergraduate accounting students. The results show significant differences between both assessment scores, with the vast majority of students overestimating their computer literacy. The authors conclude that SA cannot be relied upon as an accurate measure of entry-level computer literacy among undergraduate accounting students.

Freeman (1995) explores the accuracy of PA by examining the differences between PA and TA for final-year undergraduate accounting and finance students using different statistical tests (such as a two-sample t-test, a matched pairs t-test and a Wilcoxon matched-pairs signed-rank test). Freeman finds that staff and student average scores are not significantly different from one another, and the average presentation is scored almost identically by students and staff alike. However, there is a considerable amount of variation for individual presentations, which indicates that average marks have masked discrepancies between individual presentations.

Other studies have assessed the accuracy of both SA and PA compared to TA by means of correlation coefficients (e.g., Longhurst and Norton, 1997; Papinczak et al., 2007; Lew, Alwis and Schmidt, 2010) but have not always found it to be a very useful indicator for accuracy[6] (Fitzgerald, White and Gruppen, 2003). For example, Longhurst and Norton (1997) compare SA to TA for second-year undergraduate psychology students; they find a significant positive correlation ($r = 0.43$) between student and tutor grades. Longhurst and Norton (1997) conclude that students are reasonably accurate at grading their own essays although the magnitude of the correlation coefficient obtained

is lower than 0.50. Papinczak et al. (2007) also explore the accuracy of SA and PA relative to TA for first-year medical students using the Pearson correlation coefficient. They find that the TA correlates poorly with SA ($r = 0.31$–0.41). However, the PA appears to be slightly more accurate, with peer-average scores correlating moderately with tutor ratings (initially: $r = 0.40$); they also find that the correlation improves over time ($r = 0.60$). Lew, Alwis and Schmidt (2010) examine the accuracy of SA and PA results for first-year students enrolled in a post-secondary school institution by calculating the correlations between the scores of SA, PA and TA. They document a low correlation between TA and SA ($r = 0.23$) and a very low correlation between TA and PA ($r = −0.03$). Thus, several studies have tested the accuracy of SA and PA using different statistical techniques and reached inconsistent results.

The validity of the assessment is distinct from its accuracy and is defined as 'the extent to which any instrument measures what it is intended to measure' (Carmines and Zeller, 1991, p. 17). There are three common types of validity: (i) criterion (ii) content and (iii) construct. It is the third of these (i.e., construct validity) which is relevant to the comparison of SA/PA with TA, because it focuses on the extent to which a measure performs in accordance with theoretical expectations. Weber (1990, p.19) states that "a measure has construct validity to the extent that it is correlated with some other measures of the same construct". Therefore, a strong correlation between different measures of the same construct indicates the validity of these measures. The correlation coefficient can take a value from -1 to +1. The higher the correlation, the higher is the validity of the construct. In the context of assessment, we would aim to get a perfect

positive correlation coefficient (+1) to conclude that the variables are measuring the same thing. A number of prior studies calculate the correlation between SA and/or PA scores and TA scores to examine the validity of SA and/or PA (e.g. Cho, Schunn and Wilson, 2006; Wen and Tsai, 2008; De Grez, Valcke and Berings, 2010; Liang and Tsai, 2010) but again reach inconsistent conclusions.

2.2 Students' perception of self- and peer- assessment

Some of the existing literature also analyses student perceptions of their involvement in the self- and peer-assessment process. For example, Falchikov (1986) seek the views of 48 students who, after participating in the SA and PA of essays, feel that they have benefited from the experience because it made them "…think more, learn more, and become more critical and structured" (p. 161). Hanrahan and Isaacs (2001) also survey a cohort of 233 students who believe that they have benefited from the exercise because, apart from improving their own writing from the feedback that they received, they also gained a better understanding of the marking process, developed an empathy with tutors, saw examples of both good and bad work, and were motivated to impress their peers. Indeed, some students noted that their involvement in PA was their only opportunity to see the work of others. These authors argue that, while SA helps students to "set goals and … learn for themselves" (p. 54), PA helps them "contribute constructively in collaborative efforts" (p. 54). Stellmack et al. (2012) use PA as a "review-revise-resubmit" technique to help students improve their writing skills. These authors suggest that peer review is not only useful for those that receive feedback, but also for those that provide the feedback. This is because it allows peer assessors to compare their own work to that of others which helps them realise that, if they deem the

work they are reviewing to be "unclear and unintelligible" (p. 236), then their work might be construed in a similar way. Zundert, Sluijsmans and Merriënboer (2010) review the PA literature between 1990 and 2007 and find that, overall; there are positive student attitudes to this type of practice. Thus, students believe that both SA and PA are useful to them in a variety of different ways.

Some literature also suggests that, whilst the PA intervention is helpful and beneficial to students, the process can be time consuming, difficult and challenging (Falchikov, 1986). For example, Cheng and Warren (2006) identify a number of problems that students attributed to PA; these included: (i) too much expectation of non-native English speakers who might be required to assess the language proficiency of others; (ii) doubts about the objectivity of markers when work was not made anonymous; (iii) fears that PA marks were not accurate in comparison to TA (they felt that the latter should have a greater influence in the final mark awarded); and (iv) a lack of training to allow students to complete the PA task effectively. However, PA can be further complicated by the suggestion that not all students will react similarly (Gatfield, 2006); for example, when comparing the PA experiences of Australian and international students, Gatfield (2006) found that the latter group expressed higher levels of satisfaction. He suggested that, because the majority of these international students were Asian, their allegiance to Confucian principles supported their willingness to participate in group work and the PA process. Alternatively, Smith, Cooper and Lancaster (2010) find that PA can make students feel anxious. Thus, the literature suggests that there are a number of advantages and disadvantages associated with both SA and PA.

The current study has aims which are similar to Lindblom-Ylanne, Pihlajamaki and Kotkas (2006) who conducted a small study with 15 law students. Using essays, these authors tested the accuracy of SA/PA by comparing these marks to TA marks and explored the students' experiences of the self- and peer-assessment process. The current investigation differs from Lindblom-Ylanne, Pihlajamaki and Kotkas (2006) because it uses a larger number of students who are drawn from an accounting and finance context. Specifically it seeks to (a) establish if SA and PA are accurate and valid in comparison to TA; and (b) assess the views of these students of their SA/PA experiences. The following section explains how the paper attempts to address these two issues.

3. Method

The present study is conducted in two parts; the first part provides quantitative evidence on the accuracy and the construct validity of SA and PA for two groups of accounting and finance students, one postgraduate group in England undertaking an oral presentation and another undergraduate group in Scotland undertaking an essay writing task. The purpose is to provide further evidence on the accuracy and construct validity of SA and PA compared to TA rather than a direct comparison which is impractical to execute here due to differences between the two groups of students. All PA, SA and TA scores were produced using the same assessment criteria at each site and the marking process was fully explained to all students. The second part provides qualitative evidence about the students' perceptions of the SA and PA process which were gathered

using a questionnaire. The development of this questionnaire is also discussed in the current section.

3.1 Stage 1 – The Accuracy and Construct Validity of Self- and Peer-assessment

3.1.1 Group one

The first group of students consisted of 19 postgraduate accounting students who were studying at an English university and registered for a module entitled *Company Valuation and Performance*. Using a checklist of ten items, the students were asked to assess a verbal presentation about a case study. A one-hour seminar session was used to discuss a different case study every week for eight weeks. To adapt the case study for a team-based activity, the topic was selected, the individual or team was formed and the assessment instrument was developed (Scofield, 2005). Cases were selected from Palepu et al. (2007) and students were asked to organise themselves into groups of three.

The assessment instrument was constructed to assess the structure, content and delivery of the oral presentation and was developed based on a review of prior studies (e.g. Freeman, 1995; Langan et al., 2008; De Grez, Valcke and Berings, 2010) as well as two evaluation forms that are in use for similar purposes at two UK higher education institutions. It covered ten elements of a presentation using a five-point scale ranging from poor (1) to impressive (5). A copy of the assessment instrument used is presented in Appendix 1. Two experienced members of academic staff were asked to judge

whether the assessment instrument was appropriate to assess student performance in oral presentations (e.g. Papinczak et al., 2007). The feedback received was then used to revise and improve the assessment instrument before its use.

Of the 19 students registered for this module, 18 individuals were organised into groups of three, while the remaining student was asked to present a case study on her own.[7] Each group of students decided on the case study that they would present and a timetable for the seven presentations was developed and made available to all students. In order to provide an example for the students, a case study presentation was made by the lecturer in the first seminar. Since every student participated in the presentations, individual feedback on each of the students was collected.[8] SA, PA and TA observations took place at the same time each week using the same assessment criteria; these were collected and analysed following the presentation. Students were asked to complete the instrument privately and return their evaluations at the end of each class. PA was anonymous but, for obvious reasons, SA was not. If a student ticked more than one box for the same question, then the response to that question was considered to be invalid. One student failed to participate in the presentations; hence the final sample comprises only 18 students.

Most of the students registered for this course were from overseas[9], hence there may be an issue of potential unfamiliarity with the assessment methods used. The task was explained to them in full; but formal training was not possible due to resource constraints. Although Lew, Alwis and Schmidt (2010) find that training does not affect the accuracy of SA and PA over time, we believe that the weekly involvement of

students in PA has provided them with some sort of training. Furthermore, the assessment instrument was also fully discussed with the class prior to use, to aid students' understanding and to address any questions they might have.

Each student that took part in the study was required to self-assess themselves in the one case study that they presented and to undertake peer- assessment for the remaining cases presented by the other groups. SA, PA and TA evaluations for each of the ten elements of the presentation were collected rather than an overall mark because the latter may mask discrepancies in marking individual elements of the presentation that may cancel each other out (Orsmond, Merry and Reiling, 1996). Hence the number of total number of observations is 180.

3.1.2 Group two

The second piece of empirical work was conducted amongst 80 second year undergraduate accounting students studying a business law module in a Scottish university. PA had been used extensively during the delivery of this module[10] so all of the students were familiar with its features.

To assess the accuracy and construct validity of SA and PA, the class test for this module, which involves undertaking an essay writing task, was marked out of 25, was self-assessed, peer-assessed and tutor-assessed. Immediately after the class test was administered, two photocopies were made of each paper and the relevant student's name was noted on one copy while their matriculation number was written on the other. During the following seminar, but before the tutor marks were distributed, the seminar tutor gave a brief outline of the research project and asked the students to participate,

emphasising that they were not obliged to take part. All students present were given a copy of their own test paper, one classmate's anonymised work and a copy of the lecturer's criteria to guide their marking. Thus, the SA, PA and TA were based on the same assessment criteria that covered the content, structure and presentation of the essay. These criteria are subject specific and are available from the authors upon request. All of the photocopied test papers were collected at the end of the seminar and this exercise elicited 72 papers that were self-assessed and 80 scripts were peer-assessed.

3.2 Stage 2 Students' perceptions of self- and peer- assessment

A follow up survey was administered by the researchers at the final seminar session at each site; the students were reminded of their earlier involvement in the research project and that they were not obliged to participate. The questionnaire comprised of 12 questions. Ten questions had a Likert scale; eight of those employed a five point scale and two used a three point scale. Since the response for each question is analyzed individually, the different Likert scales employed are not considered to have an impact on the analysis or the results obtained. The final two questions were open-ended and all 12 questions provided space for students' further comments to enable a qualitative assessment of students' views about this exercise. The questions were derived from the potential advantages of SA and PA that have long been recognised in the literature (e.g., Stefani, 1994; Boud, Cohen and Sampson, 1999; Bostock, 2000; Lew, Alwis and Schmidt, 2010). The questionnaire was pilot tested, and feedback was used to revise and improve the instrument before it was used. The follow up questionnaire is presented in Appendix 2. The questionnaire was distributed to both groups of students (18 in group

one and 82 in group two) at their respective institutions which resulted in 12 and 56 responses respectively. The results were entered in an Excel spread sheet for analysis. In addition, written comments received from the students augmenting their views about this exercise were also transcribed in an Excel spread sheet. The comments were first classified according to their tone as either positive or negative and then reclassified using the general dimensions developed by Hanrahan and Isaacs (2001). Two of the co-authors have individually coded the comments, and then any discrepancies were discussed and resolved. The results have been analysed using those themes identified by Hanrahan and Isaacs (2001): (i) students found SA and PA to be difficult; (ii) SA and PA allowed students to gain a better understanding of marking; (iii) students suffered discomfort with their involvement in SA and PA; (iv) students felt the exercise was productive; (v) students considered that they benefited from exposure to other students' work; (vi) students developed any empathy with their peers; (vii) students were motivated by the exercise; and (viii) students considered there were problems with the implementation of the exercise.

4. Results and discussion

4.1 The Accuracy and Construct Validity of SA and PA

Table 1 provides the descriptive analysis of the SA, PA and TA results from the two groups of students. Panel A of Table 1 describes the results of the 18 postgraduate students from England undertaking an oral presentations whereas Panel B describes those of the 80 undergraduate students from Scotland undertaking an essay writing task. The marks of both tasks were re-scaled to 0-100 for presentation purposes only. Given

the differences between these two groups of students, the results are not pooled but rather presented seperately for each group of students; thus, no direct comparison between the two groups of students is made. In Panel A, the PA marks (mean 72.13) were the most generous compared to both the SA marks (mean 66.15) and the TA marks (mean 57.89). However, in Panel B, the SA marks (66.28) were the most generous relative to the PA (63.10) and the TA marks (56.79). The results in Table 1 also show that the lowest standard deviation is recorded for the PA and SA scores in Panel A and Panel B respectively. This finding indicates that the most generous assessment also tend to be the least variable in terms of the awarded scores. It is also interesting to note that the range of marks awarded by tutors fell in between that of the SA and PA marks for both groups of students.

<Insert table 1 here>

In accordance with Freeman (1995) and Larres , Ballantine and Whittington (2003), the Wilcoxon signed-rank test is used to examine the differences between: (a) the SA/PA marks and (b) the TA marks. The Wilcoxon signed-rank test is a non-parametric statistical test and is used because the data collected in this research are primarily ordinal. This test considers both the magnitude and the direction of the difference within the pairs and provides information about the direction of inaccuracy. Paired comparisons also provide information on the accuracy of the evaluation of individual students. This information can be used for pedagogical purposes; for example, providing feedback to students, which can help them understand their strengths and weaknesses and to reflect on their professional development. However, for the purpose of the current study, only the aggregate results are discussed.

Panel A of Table 2 shows the results of the Wilcoxon signed-rank test. It highlights the number of times students gave the same marks as the teacher (pairwise agreement), and the number of times students gave higher (overestimation) or lower (underestimation) marks than the teacher. The results show that students tended to overestimate their own marks and their peers' marks in each case, a result which is in agreement with the findings of Papinczak et al. (2007). The percentage of PA overestimation is significantly higher for group one indicating perhaps a close relationship between students due to a smaller class size. The results also show that there are significant differences between TA and those awarded through SA and PA in all cases. The results indicate that SA and PA marks were not accurate when compared to TA. These results are consistent with the findings obtained by Larres, Ballantine and Whittington (2003) and Lew, Alwis and Schmidt (2010) but contrast with findings from other studies such as those by Freeman (1995) and Evans, Leeson and Petrie (2007). These results might impede the use of SA and/ or PA for the purposes of the summative assessment. Alternatively, they may be used in conjunction with the teacher assessment for the purpose of summative assessment as long as they are scaled up or down as appropriate.

Panel B of Table 2 shows the results for the construct validity test using the Spearman correlations between PA, SA and TA at both sites. The correlation between SA and TA is significant but ranges from relatively low to moderate values (0.189 to 0.544). This result calls the construct validity of SA into question. The correlation between PA and TA is moderate and consistent for both groups of students (0.549 and

0.547). The link between the PA and TA, therefore, seems to be stronger which is consistent with findings obtained by Liang and Tsai (2010). This in turn means that both peers and teachers are roughly ranking students in a similar order consistent with findings obtained by Orsmond, Merry and Reiling (1996). The average correlation coefficient is 0.54 in all cases except for the correlation between SA and TA (0.19) for the postgraduate students. This low correlation might indicate that this particular group of students is struggling to understand their own strengths and weaknesses, and this ability needs further development (Blanch-Hartigan, 2011).

In sum, the findings show that both SA and PA are not accurate when compared to TA, which means that there are significant differences between student assessment and teacher assessment. However, PA is moderately valid, which means that both students and teachers are roughly ranking students in a similar order. The construct validity of SA is less robust, though. These results indicate that PA and SA may not be appropriate for the purpose of summative assessment. Nevertheless, this does not mean that SA and PA are not useful to students, particularly in the development of their transferable skills. The following section explores the results of the questionnaire that explored students' perceptions of the exercise.

4.2 Student Perceptions of SA and PA

The above analysis demonstrates that both SA and PA appear to be neither accurate nor completely valid compared to TA. However, this does not mean that the use of SA and/or PA is a waste of resources and the following analysis demonstrates that, despite its inaccuracy and the low to moderate levels of construct validity, students still find

these types of assessment to be useful. Table 3 summarises the results of the questionnaire distributed to students after the SA and PA exercise.

Table 3 demonstrates that the majority of students had very positive views on both SA and PA; for example, 80% of the respondents either agreed (62%) or strongly agreed (18%) that SA helped them to evaluate their own work. This result is perhaps not surprising given that, for both groups of students, the marking criteria were made very explicit and available for the students against which to assess themselves. However, the remainder of the responses is also very positive since 47% (39% and 8%) believed that this form of assessment motivated students to learn, 61% (48% and 13%) thought that it allowed them to control their learning and 68% (57% and 11%) said that it helped them to develop a valuable skill for their future professional career. However, only 29% of the respondents stated that they were comfortable assessing their own work which would indicate that, whilst they found it a positive experience, they had concerns regarding their own objectivity.

The student views relating to PA were similarly positive; for example, 62% of respondents either agreed (52%) or strongly agreed (10%) that PA helped them to evaluate their own work. Although this is a slight decrease on the 80% who believed that SA helped them to assess their own work, it still suggests that, when reviewing the work of their peers, a large proportion of these students were able to make comparisons between these two pieces of work. Again, the explicit nature and availability of marking criteria might help explain why 77% (67% and 10%) of respondents believed that PA

helped them to evaluate the work of others; this is supported by the 67% (49% and 18%) of respondents who believed that PA helped them to better understand the marks awarded by the tutor. Similar to the 47% of students who believed that SA motivated them to learn, 43% (41% and 2%) of respondents believed that the PA also motivated them to learn. Interestingly, more students (56%) were comfortable assessing the work of others than they were assessing their own work which may indicate that they were more comfortable working with increased levels of objectivity. However, despite these positive comments the majority of students (59%) agreed that they would still prefer their summative assessment to be marked by the teacher.

Comments written by the students augmenting the views they had given in the survey are analyzed and discussed in the remainder of this section. In the current study, we received 163 useful comments that are classified as positive or negative in tone. Then these 163 comments are reclassified using the eight general dimensions developed by Hanrahan and Isaacs (2001).

<Insert table 4 about here>

Table 4 shows a summary of students' comments regarding their views of the SA and PA. It shows that 106 comments were perceived as positive, while 57 comments were perceived as negative in tone. Consistent with Hanrahan and Isaacs (2001), the negative comments were classified under three themes: (i) difficult, (ii) discomfort, and (iii) problems with implementation, while the positive comments were classified under five themes: (i) better understanding, (ii) productive, (iii) motivation, (iv) empathy and finally (iv) others' work. Specific comments that were considered to indicate that

students found SA and PA difficult included '*I am unsure of what is correct and where I should allocate marks*' while another student found they '*cannot judge themselves. When I judge my work sometimes I think it is perfect*' which contrasted with another respondent's feeling that '*I tend to be harder on myself*'. The 14 comments seem to indicate that certain students experienced discomfort during the SA and PA processes included one student who was '*often scared of returning poor feedback*'. The most common negative view of SA and PA exercises relates to problems with their implementation particularly where certain students failed to fully engage with the process. One student commented '*A lot of the time classmates had not submitted any work and throughout the weeks I did not mark any decently done essay*'. Frustration was also evident where students submitting work failed to get the feedback they expected as one respondent noted '*I didn't always receive feedback and when I did it was too vague to be of any use*'.

Positive comments include responses that showed students felt they had gained a better understanding of marking through the assessment exercise included '*It helps when it comes to understanding how essays are answered and how examiners want us to answer them*' and '*it's the best way to learn from your mistakes and improve*'. Others felt the exercise encouraged them to be productive and would be useful in the workplace '*I will be expected to carry out CPD* [Continuing Professional Development] *to improve my knowledge and skills and through self-assessment I will be able to do it*'. Several students indicated that the SA and PA exercises had helped their motivation and 19 comments were made to support this. One student admitted '*I wouldn't have written the essays otherwise*' while another felt '*You can see how well others are doing which*

can motivate you to do well. Also the feedback you receive can be motivating'. Also, the students appeared to develop some empathy with their peers and some found it difficult to be too critical with one comment reading '*You have to criticise yourself before others'* and another saying '*I prefer others marking my work as I am afraid I will be either too harsh or biased'.* Four comments related to thoughts regarding the opportunity of seeing another student's work and it seemed to suggest that this was helpful but important that the work was anonymised. Overall, the survey results indicate that students have positive perceptions of both SA and PA, although some difficulties are also documented.

Asked for their views on how SA and PA might be improved, most responses referred back to the problem of certain students' lack of engagement with the process with comments such as making sure everybody participates by submitting a meaningful piece of work and providing quality feedback rather than just noting a '*random grade'.* It would be useful when this exercise is repeated to examine ways in which each student's effort in SA and PA could be quantified and reflected in the module result.

5. Concluding remarks

This paper contributes to the growing body of research on SA and PA by providing further insights from the context of a rarely investigated discipline, i.e. accounting and finance. In particular, the paper attempts to answer the following questions: (i) Are SA and PA accurate measures of student performance? (ii) Are SA and PA valid measures of student performance? (iii) How do students perceive SA and PA?

The results show statistically significant differences between SA, PA and TA which indicate that SA and PA do not provide accurate measures of student performance when compared to TA. In addition, the construct validity of SA ranges from low to moderate. In contrast, PA seems to be moderately valid and robust to the different conditions employed in the current study. These results indicate that students have some ability to evaluate their own performance as well as their peers' performances, however this ability is limited. These results suggest that the use of SA and PA might not be appropriate for the purpose of summative assessment because they do not provide either an accurate or a completely valid measure of student performance. However, this should not prevent their implementation for the purpose of formative assessment, since both SA and PA are considered valuable by students in their learning process. Teachers may consider including self- and peer- assessment exercises for the purpose of formative assessment on a regular basis when designing and staging their courses. This in turn would give students continuous feedback on their learning process and perhaps improve their performance. The outcomes of this investigation could also be useful in the implementation of an education policy that is designed to teach students how to implement the standards employed by their teachers.

In the current study, the assessment instrument was intentionally designed by the teacher to cover specific aspects of student performance. Although Langan et al. (2008) found that student participation in the development of assessment criteria has no significant effects on the results of the SA, if students take a more active role in developing the assessment instrument, this may motivate them to generate more reasonable SA and PA evaluations. In addition, the current study investigated students'

perceptions of SA and PA; a future study might alternatively investigate factors that affect the students' evaluation. Such investigation can help us understand why students overmark or under mark themselves and/or their peers and may therefore help us to reduce or eliminate discrepancies. Going forward, it would be useful if some system could also be devised that motivates more students to become involved in the SA and PA exercises. If this is developed and implemented it would be useful to repeat this study to see if there is much change in students' views of the value of SA and PA. Finally, while we relied on the results obtained from two cohorts of students to draw a conclusion about the accuracy and construct validity of SA, PA and TA in the accounting and finance field, the size of the sample, in particular for the postgraduate students, is small. Future research may replicate this exercise for larger samples of students to reach firm conclusions. In addition, a direct comparison between the two groups of students was not practical in the current study given the differences between them; future research may replicate the analysis for similar groups of students undertaking different tasks to check the sensitivity of the results to different tasks.

Acknowledgement

Learning & Teaching Symposium at Brunel University 2011 for their helpful comments.

Notes

---

[1]The main object of formative assessment is to provide learners with feedback on how they are performing during a programme of study thereby helping them to learn more effectively. It does not normally count towards a final grade nor is it normally used to determine whether the learner will be allowed to progress to a later stage of a course. It is, however, sometimes used to permit entry to an examination – the class certificate or 'duly performed' approach (Ellington 1996).

[2]Summative assessment is normally conducted at the end of study to establish what the learner has achieved. It differs from formative assessment in that it usually counts towards a final grade or is used to determine whether the learner is allowed to progress through the course (Ellington 1996).

[3] Orpen (1994) provides an example of tests for predictive validity in the context of peer-assessment. Individuals are asked to predict a future achievement for their peers, in this case, the future examination performance of their peers. Then, the correlation coefficients between these predictions and actual outcomes are calculated to determine the predictive validity of the peer evaluations. However, the predictive validity is not one of the tests of interest to the current study.

[4] Johnson and Smith's (1997) study is one of the few to examine peer assessment in the accounting and finance domain. However, they study the concurrent validity of peer evaluations which is not the subject of the current study. In their research, peer assessment is the score awarded to each student by his or her team members using a proposed instrument. To examine the concurrent validity of peer evaluation, the Spearman rank correlation coefficient is calculated between peer evaluations and individual objective scores based on a quiz. A significant correlation coefficient of 0.46 is obtained between peer scores using the proposed evaluation instrument and the individual score, which led Johnson and Smith to conclude that their proposed instrument produced reasonably valid peer scores.

[5] This is also true of reliability which refers to the extent to which an experiment or test yields similar results on repeated trials (Carmines and Zeller, 1991). Indeed, the Chartered Institute of Educational Assessors (2011) suggests that an assessment is reliable where a candidate scores the same mark irrespective of who the assessor is. In the context of SA and PA, most of the existing literature uses the inter-marker reliability as a measure for reliability (Stefani, 1994; Evans, Leeson and Petrie., 2007; Papinczak et al., 2007). Topping (1998) suggests that, to provide a true measure of reliability, markers must be at the same level of education, training and professionalism, otherwise a measure of validity is obtained.

[6] This is because, with a correlation coefficient one does not know how different the student assessment and the teacher assessment are or which is higher than the other, only whether they move together in the same or opposite direction or not moving together at all (Blanch-Hartigan, 2011).

[7] Because the teaching plan of this course was set up to analyse one case study every week, this resulted in one student handling a case study by herself. Hence her work was minimised by requiring her to present a case study only where the full answer was available.

[8] Submissions for peer assessment were invited the week after the first group's presentation, which resulted in very few responses. Hence it was decided to drop the peer review observations for the students who formed the first group. Afterwards, evaluation forms were collected at the end of each class.

[9] Another potential issue is the possibility of students being unfamiliar with case studies. Hence an illustration of how to handle a case study was provided by the lecturer at the first seminar. Students were encouraged to discuss with the lecturer prior to presentation any aspects of the case study that they were required to analyse and present. During discussion of each case study, the lecturer also demonstrated to

the class the required steps for analysis. Furthermore, when necessary, the correct answer to the case study was made available to all students after presentation. However, future research might replicate the analysis controlling for the effect of the assessment task prior to reaching firm conclusions.

[10] Each week the students were required to upload a written piece of work (similar to an exam-type question) onto the university virtual learning environment. Computer software was used to anonymise the work and then to distribute two pieces of their peers' work to each student which they were required to mark in accordance with marking criteria that had been developed by the module tutors. Once the work was marked and written feedback supplied, the software then returned the feedback to the author. Thus, each student marked two pieces of their peers' work and received two pieces of peer-feedback. This exercise was undertaken in the university's computer labs.

References

Blanch-Hartigan, D. (2011) Medical students' self-assessment of performance: results from three meta-analyses, *Patient Education and Counseling,* 84, pp. 3-9.

Bostock, S. (2000) Student peer assessment. The Higher Education Academy. Available on line at: http://www.reading.ac.uk/web/FILES/engageinassessment/Student_peer_assessment_-_Stephen_Bostock.pdf [last accessed 9/11/2013].

Boud, D. and Lublin, J.R. (1983) Self-assessment in professional education. Report to the Commonwealth Research and Development Committee, Tertiary Education Research Centre, University of New South Wales.

Boud, D., Cohen, R. and Sampson, J. (1999) Peer Learning and Assessment. *Assessment and Evaluation in Higher Education*, 24, pp. 413-426.

Dochy, F., Segers, M. and Sluijsmans, D.M.A. (1999) The use of self-, peer and co-assessment in higher education: a review, *Studies in Higher Education,* 24 (3), pp. 331-350.

Carmines, E.G. and Zeller, R.A. (1991) Reliability *and Validity Assessment.* London: Sage.

Chartered Institute of Educational Assessors (CIEA), available at: http://www.ciea.org.uk/knowledge_centre/articles_speeches/general_articles/assessment_relia bility.aspx [last accessed 9/10/11].

Cheng, W. and Warren, M. (2006) Having second thoughts: Student perceptions before and after a peer assessment exercise, *Studies in Higher Education*, 22 (2), pp. 233-239.

Cho, K., Schunn, C.D. and Wilson, R.W. (2006) Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives, *Journal of Educational Psychology* , 98 (4), pp. 891–901.

De Grez, Valcke, and Berings (2010) Peer assessment of oral presentation skills, *Procedia – Social and Behavioral Sciences,* 2 (2), pp. 1776–1780.

Ellington, H. (1996) Assessing student performance: a course booklet for the postgraduate certificate in tertiary-level teaching. The Robert Gordon University. http://www.nalanda.nitc.ac.in/misc/general/ciced/Ch21.html

Evans, A.W., Leeson, R.M.A. and Petrie, A. (2007) Reliability of peer and self-assessment scores compared with trainers' scores following third molar surgery, *Medical Education,* 41, pp. 866–872.

Falchikov, N. (1986) Product comparisons and process benefits of collaborative peer- and self-assessments, *Assessment and Evaluation in Higher Education,* 11, pp. 146-166.

Falchikov, N. and Boud, D. (1989) Student self-assessment in higher education: a meta analysis. *Review of Educational Research,* 59 (4), pp. 395–430.

Falchikov, N. and Goldfinch, J. (2000) Student peer assessment in higher education: a meta analysis comparing peer and teacher marks. *Review of Educational Research,* 70 (3), pp. 287–322.

Fitzgerald, J.T., White, C.B. and Gruppen, L.D. (2003) A longitudinal study of self-assessment accuracy. *Medical Education,* 37, pp. 645–649.

Freeman, M. (1995) Peer assessment by groups of group work. *Assessment & Evaluation in Higher Education,* 20 (3), pp. 289–300.

Gatfield, T. (2006) Examining student satisfaction with group projects and peer assessment, *Assessment & Evaluation in Higher Education,* 24 (4), pp. 365-377.

Hanrahan, S. and Isaacs, G. (2001) Assessing self- and peer-assessment: the students' views, *Higher Education Research &Development,* 20 (1), pp. 53-70.

International vocabulary of basic and general terms in metrology (VIM), The International Vocabulary of Basic and General Terms in Metrology, 2006, http://www.nist.gov/pml/div688/grp40/upload/International-Vocabulary-of-Metrology.pdf [last accessed 28//5/ 2013].

Johnson, C.B. and Smith, F.I. (1997) Assessment of a complex peer evaluation instrument for team learning and group processes, *Accounting Education,* 2 (1), pp. 21–40.

Langan, A.M., Shuker, D.M., Cullen, W.R., Penney, D. and Preziosi, R.F. (2008) Relationships between student characteristics and self-, peer and tutor evaluations of oral presentations. *Assessment & Evaluation in Higher Education,* 33 (2), pp. 179–190.

Larres, P.M., Ballantine, J.A. and Whittington, M. (2003) Evaluating the validity of self-assessment: measuring computer literacy among entry-level undergraduates within accounting degree programmes at two UK universities, *Accounting Education,* 12 (2), pp. 97–112.

Lew, M.D.N., Alwis, W.A.M. and Schmidt, H.G. (2010) Accuracy of students' self-assessment and their beliefs about its utility, *Assessment & Evaluation in Higher Education,* 35 (2), pp. 135–156.

Liang, J.C., and Tsai, C.C. (2010) Learning through science writing via online peer assessment in a college biology course, *The Internet and Higher Education,* 13 (4), pp. 242–247.

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

 Longhurst, N., and Norton, L.S. (1997) Self-assessment in coursework essays, *Studies in Educational Evaluation,* 23 (4), pp. 319–330.

Lynn, D. J, Holzer, C., and O'Neill, P. (2006) Relationships between self-assessment skills, test performance, and demographic variables in psychiatry residents, Advances in *Health Sciences Education*, 11, pp.51–60.

Matsuno, S. (2009) Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms, *Language Testing,* 26, pp. 75–100.

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

Orpen, C. (1994) Perceived similarity: Its effect on the accuracy of peer evaluations among university students, *International Journal of Educational Management,* 8, pp. 4-6.

Orsmond, P., Merry, S. and Reiling, K. (1996) The importance of marking criteria in the use of peer assessment,Assessment & Evaluation in Higher Education, 21 (3), pp. 239–250 .

Palepu, K. G., Healy, P. M., Bernard, V. L. and Peek, E. (2007) Business Analysis and Valuation: IFRS Edition –Text and Cases. Cengage Learning EMEA, United Kingdom.

Papinczak, T., Young, L., Groves, M. and Haynes, M. (2007) An analysis of peer, self, and tutor assessment in problem-based learning tutorials, *Medical Teacher* , 29 (5), pp. 122–132.

Regehr, G. and Eva, K. (2006) Self-assessment, self-direction, and the self-regulating professional, *Clinical Orthopaedics and Related Research*, 449, pp.34–8.

Scofield, B.W. (2005) Adapting cases for a team approach, *Journal of Accounting Education,* 23 (4), pp. 248–263.

Smith, H., Cooper, A. and Lancaster, L. (2010) Improving the quality of undergraduate peer assessment: a case for student and staff development, *Innovations in Education and Teaching International*, 39 (1), pp. 71-81.

Stefani, L.A.J. (1994) Peer, self and tutor assessment: relative reliabilities. *Studies in Higher Education,* 19 (1), pp. 69–75.

Stellmack, M., Keenan N., Sandidge R., Sippl A. and Konheim-Kalkstein, Y. (2012) Review, revise and resubmit: the effects of self-critique, peer review and instructor feedback on student writing, *Teaching of Psychology,* 39 (4), pp. 235-244.

Topping, K.J. (1998) Peer assessment between students in college and university, *Review of Educational Research,* 68 (3), pp. 249–267.

Weber, R.P. (1990) *Basic Content Analysis*. 2nd ed. London: Sage.

Wen, M.L. and Tsai, C. (2008) Online peer assessment in an inservice science and mathematics teacher education course, *Teaching in Higher Education,* 13 (1), pp. 55–67.

Zundert, M., Sluijsmans, D. and Merriënboer, J. (2010) Effective peer assessment process: research findings and future directions, *Learning and Instruction*, 20, pp. 270-279.

Table 1. Descriptive statistics

	N	Min.	Max.	Mean	STD
<u>Panel A: Group one</u>					
Self-assessment (SA)	179	20	100	66.15	19.67
Peer-assessment (PA)	150	40	80	72.13	10.14
Tutor assessment (TA)	180	20	100	57.89	17.50
<u>Panel B: Group two</u>					
Self-assessment (SA)	72	32	100	66.28	14.55
Peer-assessment (PA)	80	20	100	63.10	17.16
Tutor assessment (TA)	96	24	92	56.79	16.47

Table 2. Tests of the accuracy and construct validity of SA and PA

Panel A: Test of accuracy: Wilcoxon signed-rank test.

Group one		N	%	Mean Rank	Z
SA vs. TA	SA<TA	37	0.207	67.500	-4.274**
	SA>TA	93	0.520	64.704	(0.000)
	SA=TA	49	0.274		
	Total	179			
PA vs. TA	PA<TA	2	0.013	49.000	-9.573**
	PA>TA	112	0.747	57.652	(0.000)
	PA=TA	36	0.240		
	Total	150			
Group two					
SA vs. TA	SA<TA	21	0.292	25.929	-3.725**
	SA>TA	46	0.639	37.685	(0.000)
	SA=TA	5	0.069		
	Total	72			
PA vs. TA	PA<TA	25	0.313	32.400	-2.538**
	PA>TA	45	0.563	37.222	(0.011)
	PA=TA	10	0.125		
	Total	80			

Panel B: Test of construct validity: Spearman correlation

		Tutor assessment	
		Group one	Group two
Self- assessment	Correlation Coefficient	0.189**	0.544**
	Sig. (2-tailed)	(0.011)	(0.000)
	N	179	72
Peer- assessment	Correlation Coefficient	0.549**	0.547**
	Sig. (2-tailed)	(0.000)	(0.000)
	N	150	80

SA: self-assessment; TA: tutor assessment; PA: peer- assessment; Values in parentheses are probabilities of significance.   Significant at 1% level (two-tailed).

Table 3. Results of the Self- and Peer- assessment evaluation Questionnaire

	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree	Average Rating
Panel A: Self-assessment						
Develops ability to evaluate own work	4	4	12	62	18	3.90
Motivates student to learn	6	16	31	39	8	3.30
Allows student to control learning	5	10	24	48	13	3.56
Develops professional skills	3	3	26	57	11	3.73
Student comfortable assessing own work	31		40	29		1.97
Panel B: Peer- assessment						
Develops ability to evaluate own work	0	14	24	52	10	3.65
Develops ability to evaluate others' work	3	5	15	67	10	3.80
Helps understand tutor marks	3	9	21	49	18	3.75
Motivates student to learn	7	17	33	41	2	3.14
Students comfortable assessing others' work	14		30	56		2.47

Note: Table 3 reports the % student responses to a self- and peer-assessment evaluation questionnaire. Eight questions were measured on a five point Likert scale and two were measured on a three point Likert scale.

Table 4. A summary of students' comments regarding their views about the SA and PA

Negative comments		Positive comments	
Difficult	16	Better understanding	59
Discomfort	14	Productive	17
Problems with implementation	27	Motivation	19
Total	57	Empathy	7
		Others' work	4
		Total	106

Appendix 1. The assessment instrument for the oral presentation

	Impressive (5)	Very good (4)	Good (3)	Fair (2)	Poor (1)
1. Standard and quality of the content					
Structure – logical flow and cohesiveness					
Amount and scope of content relative to the time allowed					
Familiarity with the topic and ability to emphasise key points					
Ability to deal with questions and challenges from the audience					
Summations/conclusions Ability to highlight and establish links between the key issues addressed in the discussion					
2. Style and delivery					
Ability to keep an eye contact and rapport with the audience					
Voice – audibility (Could you hear the presenter clearly throughout?)					
Rate of speech (Was the pace of the speech or flow of ideas adequate?)					
Use of visual aids (Was there a suitable amount? Were they easy to read? Did they effectively support the verbal delivery? Did the presenter use them competently?)					
Body language and gesture (Was the presenter's posture upright and confident? Did the presenter's movement and gesture enhance, not distract from, what he or she was saying?)					

Appendix 2. The follow up questionnaire

This survey is a part of a research project titled: "self- and peer-assessment: evidence from the accounting and finance discipline". The purpose is to examine the accuracy and validity of peer- and self- assessments in the discipline of accounting and finance. It also examines how students perceive self and peer assessments.

Your reply will be treated as confidential. I guarantee anonymity. Copies of the questionnaire results will be available upon request. Thank you very much for your cooperation.

For any information, please do not hesitate to contact us.

Yours sincerely,

Student ID (optional) ----------------

1. Self- assessment develops my abilities to assess and evaluate my own work.

   a) Strongly disagree
   b) Disagree
   c) Neither agree nor disagree
   d) Agree
   e) Strongly agree

Because…

2. Self- assessment is a valuable skill for my future profession.

   a) Strongly disagree
   b) Disagree
   c) Neither agree nor disagree
   d) Agree
   e) Strongly agree

Because…

3. Self- assessment helps me to take control of my learning.

   a) Strongly disagree
   b) Disagree
   c) Neither agree nor disagree
   d) Agree
   e) Strongly agree

Because…

4. Self- assessment improved my motivation to learn.

    a) Strongly disagree
    b) Disagree
    c) Neither agree nor disagree
    d) Agree
    e) Strongly agree

Because…

5. How comfortable are you when assessing your own work?

    a) Uncomfortable
    b) Neither comfortable nor uncomfortable
    c) Comfortable

Because…

6. Peer- assessment (assessing the work of my classmates) develops my abilities to assess and evaluate others' work.

    a) Strongly disagree
    b) Disagree
    c) Neither agree nor disagree
    d) Agree
    e) Strongly agree

Because…

7. Peer- assessment develops my abilities to assess and evaluate my own work

    a) Strongly disagree
    b) Disagree
    c) Neither agree nor disagree
    d) Agree
    e) Strongly agree

Because…

8. Peer- assessment helps me understand the marks from my tutor.

- a) Strongly disagree
- b) Disagree
- c) Neither agree nor disagree
- d) Agree
- e) Strongly agree

Because…

9. Peer- assessment improved my motivation to learn.

- a) Strongly disagree
- b) Disagree
- c) Neither agree nor disagree
- d) Agree
- e) Strongly agree

Because…

10. How comfortable are you when assessing your classmates' work?

- a) Uncomfortable
- b) Neither comfortable nor uncomfortable
- c) Comfortable

Further comments

11. How would you prefer your summative assessment (an assessment that counts towards your final mark) to be evaluated? Please choose from the following:

a) Self- assessed
b) Peer- assessed
c) Assessed by the lecturer /tutor
d) A combination of a and b
e) A combination of a and c
f) A combination of b and c
g) A combination of a, b and c

Further comments:

12. How do you think we can improve self- and peer- assessment? Please choose from the following (you might choose more than one option):

a) Provide training
b) Others, please specify?

Further comments