THE RELIABILITY OF PEER ASSESSMENT IN A FINAL–YEAR INFORMATION SYSTEMS COURSE

S.J Kruger
School of Accountancy
Stellenbosch University
Stellenbosch, South Africa

ABSTRACT

The perceived lack of reliability is one aspect of peer assessment that makes educators reluctant to use it in their teaching. This study was done among third–year information systems students who had to assess two assignments of their peers by using rubrics. The rubrics differed in the sense that the one was more detailed requiring less subjective judgement while the other one was less detailed and required more subjective judgement from assessors. To determine the reliability of the peer assessment process, all the assignments of the participants in the study were re–assessed by the lecturer and the marks compared. While the reliability of the peer assessment of both projects was acceptable, the assessment of the assignment requiring less judgement was more reliable compared to the assignment requiring more subjective judgement. A survey done among the participants revealed that they were happy with their marks and with the effort put in by their peers in assessing them. A significant number of students believed their peers awarded them higher marks for both assignments; however, this was only found to be the case for the assignment requiring more subjective judgment. It was also found that while the ratings of one peer rater resulted in reliable marks when the detailed rubric was used, it was not the case for the rubric requiring more subjective judgement. This study provides motivation for educators to use peer assessment as fears about the reliability thereof is unwarranted provided that properly designed rubrics are used, and students are adequately guided through the process.

Key Words: peer assessment, reliability, rubrics, accounting education

INTRODUCTION

> "Peer assessment involves students in making judgements of their peers' work. It presents learners with a complex challenge, typically requiring them to reference peers' work against assessment criteria" (Jones and Alcock 2014).

Peer assessment offers potential pedagogic benefits, such as improving feedback (Nicol, Thomson and Breslin 2014, Luaces, Díez and Bahamonde 2018, Lin 2018), improving learning performance (Lundstrom and Baker 2009; Sudhakar, Tyler and Wakefield 2016), improving

writing skills (Ramon–Casas et al. 2019), and developing professional competencies (Sridharan, Muttakin & Mihret 2018; Barac, Kirstein and Kunz 2021). It can also be of value in making students responsible for their own learning (McDonald and Boud 2003, Ortega–Ruipérez and Correa–Gorospe 2024).

Peer evaluation and appraisal of colleagues' work constitute a fundamental component within the sphere of professional accounting practice (Westermann, Bedard and Earley 2015) as well as in other fields of employment (Sridharan, Tai and Boud 2019). Surprisingly, the accounting education literature appears to offer limited attention to this critical aspect (Barac et al 2021) A scant number of articles within the domain of accounting education advocate for peer assessment as a viable pedagogical approach (e.g., Ballantine and Larres 2007; McIsaac and Sepe 1996; Riley and Simons 2013; Hassan, Fox and Hannah 2014; Phillips 2016; Matherly and Burney 2009; Dale–Jones, Hancock and Willey 2013).

In contrast, the concept of peer assessment has garnered significant interest among educators and researchers in numerous other academic disciplines within higher education, spanning the physical sciences, social sciences, and humanities, as well as various professions such as computer science, engineering, and nursing (for comprehensive reviews, see Topping 1998; Falchikov and Goldfinch 2000; Van Zundert, Sluijsmans and Van Merriënboer 2010) The research and developments in these diverse fields have yielded valuable pedagogical tools and empirical insights that have the potential to enlighten and aid accounting educators in the effective integration of peer assessment practices (Phillips 2016).

Despite its well–documented advantages, the utilisation of self– and peer assessment for summative evaluation purposes has been recognized as a contentious practice, primarily due to apprehensions related to its validity and reliability (Sung et al 2010; Li et al 2016; Liu and Carless 2006). Sullivan (2011,119) defines reliability as follows: Reliability refers to whether an assessment instrument gives the same results each time it is used in the same setting with the same type of subjects. Reliability essentially means consistent or dependable results."

For peer assessment to be reliable, marks that would have been awarded by lecturers should be in line with the marks awarded by the peers of students (Jones and Alcock 2014). Comparing the grades awarded by peers and lecturers should provide clarity about the reliability of this assessment tool and dispel unfounded beliefs (Saito and Fujita 2004).

RESEARCH PROBLEM

The research problem that was identified is whether peer assessment done by accountancy students is sufficiently reliable to be used for assessment in an information systems course.

RESEARCH OBJECTIVES

The first objective of the study was to determine whether the marks given by students were not significantly different from those awarded by the lecturer of the course. It was however also important to determine whether students believed the marks awarded by their peers were in line with what they would have expected to be awarded by the lecturer (McGarr and Clifford 2013). Therefore, the second objective of the study was to determine what the perceptions of students were regarding the reliability of assessments done by their peers.

RESEARCH METHODOLOGY

The students did two projects individually as part of the summative assessment for the third year information systems course. For the first project students had to write a report where management accounting principles had to be applied (MA–R) and for the second project audit working articles had to be compiled (AU–WP). Both projects required students to use management accounting and auditing principles as well as to apply computer application skills including Excel and Word. After completion, two students were randomly assigned to mark one other student's project. Projects were submitted electronically, and the marks were awarded based on a rubric which was completed online on the university's web–based teaching and learning support system. The marking rubric for AU–WP consisted of 28 questions with very specific criteria which had to be evaluated, e.g. "Did the student test for obsolete stock?" The marking rubric for MA–R consisted of 7 questions which totalled up to 10 marks. Six questions required very specific criteria to be evaluated but the remaining question required the assessor to give a mark out of 3 for "General Impression". Marks awarded requiring subjective judgement therefore weighed more in the rubric for MA–R compared to the rubric for AU–WP.

The peer assessment was not done anonymously which meant that the students knew the name of the student they were assessing, and the students being assessed knew who assessed them. This process was completed under the guidance of a lecturer during class time. The average mark of the two assessors was allocated as the mark achieved by the student for their project.

The MA–R project contributed 8 per cent towards the predicate mark while the TR–W project contributed 17 per cent towards the predicate mark. The rest of the predicate mark was made up of an assignment which contributed 25 per cent, and a term test which contributed 50 per cent of the mark. The credits allocated for the peer–assessed projects were therefor sufficient to ensure students put in sufficient effort into the assignments without lecturers

feeling that they need to devote significant attention to grading details (Matherly and Burney 2009). Students had to achieve a predicate mark of 40 per cent to be allowed access to the exam. Final marks were calculated by giving the exam and predicate mark equal weighting. To pass the course, students had to achieve a final mark or an exam mark of 50 per cent or more.

Students were invited to complete a questionnaire about their perception of the peer assessment process. These responses are summarised in Table 3.

The lecturer presenting the course assessed all the projects of the participants using the same respective rubrics. A comparison of the marks awarded by the students and lecturer is presented in Tables 1 and 2.

RESEARCH QUESTIONS

To determine whether the peer assessment process was reliable, the following research questions were posed:

- Were the marks awarded by students substantially the same as those awarded by the lecturer for each of the assignments?
- Did students believe that the marks awarded by their peers were in line with what the lecturers would have awarded them?
- Was there a significant difference in the reliability of the peer assessment between the two assignments, given that more subjective judgement was required when assessing MA–R compared to AU–WP?
- Does having the assignments assessed by two students (and then using the average mark of the two) lead to more reliable results compared to just using the assessment of one student?

A study of literature on the reliability of peer assessment informed the decision to conduct an empirical study, by means of a questionnaire and comparison of the marks awarded by the lecturer and students.

LITERATURE REVIEW

Summative and formative assessments are generally considered to be the responsibility of lecturers (Nicol and MacFarlane–Dick 2006) and although a number of meta–analyses concluded that the reliability of peer assessment is adequate in a wide variety of applications (Falchikov and Goldfinch 2000, Topping 1998; Li et al. 2016), concerns remain around the capacity of students to consistently and accurately evaluate the work of their peers, a task contingent upon several antecedent factors. These factors encompass students' comprehension

of criteria pertaining to quality, standards, and performance expectations (O'donovan, Price and Rust 2004), as well as their aptitude for evaluative discernment (Nicol, Thomson and Breslin 2014). It is worth noting that the issues plaguing peer assessment share similarities with broader challenges concerning the judgment rendered by evaluators (Bloxham et al. 2016). Studies comparing the assessment done by educators with that of students take the educator's assessment as correct, which is not necessarily true (Langan et al. 2005). Notably, the distinction may lie in concerns related to equitable, unbiased, and candid evaluation practices (Willey and Gardner 2010). Although educators typically lack substantial incentives to inaccurately evaluate students, social obligations, such as peer pressure, among students may introduce additional distortions in grading. Therefore, adopting strategies that mitigate these social pressures in the conduct of peer assessment may enhance its acceptability, both for students and educators. The implementation of anonymous grading has been put forward as a means to enhance accuracy, as it assures students that their identities will not be disclosed when indicating areas of underperformance in their peers' work (Lin 2018; Sridharan, Tai and Boud 2019), although a meta–analysis by Li et al. (2016) found that non–anonymous peer assessment was more reliable. This could be due to improved accountability when the identity of the assessors is known by their peers.

Extensive literature reviews conducted by Boud, Cohen and Sampson 1999; Gielen et al (2011) and Kollar and Fischer (2010) have revealed a set of instructional design principles that could enhance students' performance in peer assessment. One pivotal design principle of direct theoretical relevance to the current investigation suggests that peer assessment should incorporate predefined assessment criteria (Orsmond, Merry and Reiling 1996; Topping 1998). It is imperative that assessment criteria are made clear in an understandable and comprehensive manner, as this facilitates the establishment of a shared understanding among educators and students regarding the evaluative aspects (Orsmond, Merry and Reiling 2000). The alignment of perspectives between instructors and students holds significance in ensuring the validity of assessment outcomes. Consequently, research designs that introduce students to assessment criteria, either through structured training (e.g., Cho, Schunn and Wilson 2006) or by involving them in the formulation of these criteria (e.g., Orsmond, Merry and Reiling 2000, Van Zundert, Sluijsmans and Van Merriënboer 2010; Kilgour et al. 2020) tend to yield results that are not only more valid but also more reliable (Falchikov and Goldfinch 2000).

Rubrics

Subjectivity in grading can be reduced by a well–designed rubric (Matherly and Burney 2009). A rubric is defined as "a document with scoring criteria and detailed descriptions that specify levels of performance" (Zhang, Li and Zhang 2024). Rubrics have been widely employed to improve the peer assessment process (Taylor, Kisby and Reedy 2023). Over and above being used as scoring tools to ensure consistent grading for different cohorts of students (Petkov et al. 2008), rubrics also have other benefits, such as improving academic performance and self–efficacy (Panadero et al 2023; Karaman 2024), improving feedback (Cockett and Jackson 2018), reducing the anxiety of students tasked with assessing their peers (Nawas 2020, Taylor, Kisby and Reedy 2023), and helping students understand standards they need to meet (To, Panadero and Carless 2022). Alternative methods of assessment, including peer assessment, should also meet the demands of equity, validity and reliability, and when a rubric is used it should result in similar ratings if used by different assessors (Andrade 2005).

One method to improve the effectiveness of a rubric is to co–create it with students. Yan (2024) found that when students were made part of establishing assessment criteria used in a rubric, the quality of feedback improved. However, other researchers found that students who did not participate in the co–construction of a rubric used it just as much and achieved the same performance gain as students who participated in the construction of the rubric (Zhang, Li and Zhang 2024). Guiding students on how to use rubrics is also important in order to achieve an acceptable level of accuracy (Charamba and Dlamini–Nxumalo 2022).

While a perception might exist that rubrics are only reliable when specific answers are required, acceptable reliability can be achieved even when descriptive questions are evaluated without clear assessment criteria (Son and Ha 2024). Where more subjective judgements are required, the accuracy of peer assessments can be improved by providing comparative examples to students as to which efforts are considered good, average or bad, and by specifically training them to make better judgments (Jones and Alcock 2014; Berry, Huang and Exter 2023).

Some authors have criticised the use of rubrics. They argue that rubrics are not suited to measuring complex thinking skills, and that they limit creative responses and create the false impression that the grading process is completely objective (Bennett 2016). Apart from employing rubrics to improve the reliability of peer assessment, the use of artificial intelligence where probabilistic and text–based models were employed also improved the accuracy of peer grading (Darvishi et al. 2022) .

Number of assessors

Another factor to consider which could influence the reliability of the mark awarded to a student by their peers is the number of assessors whose marks are considered to calculate the final mark awarded to the student. It can be argued that the risk of a substantially incorrect mark awarded to the student is reduced if the average mark of more than one assessor is used compared to using the mark awarded by only one assessor.

Some researchers found that 3–4 assessors were sufficient to provide reliable results when peer assessment of individual work is performed (Cho, Schunn and Wilson 2006, Sung et al. 2010). While it is generally acknowledged that single ratings are less reliable than multiple ones, Fagot (1991) as well as Falchikov and Goldfinch (2000) found that ratings produced by a very large number of assessors were less similar to those produced by the teacher compared to a smaller number of raters and that single reviewers were as reliable as multiple reviewers. This could be due to students taking less care when marking the work of their peers if they know many others would be tasked to do the same.

SUBJECTS, DATA COLLECTION AND DATA ANALYSIS

After obtaining approval from the ethics committee, an e–mail was sent to all course participants inviting them to complete the online questionnaire. Altogether 111 of the 496 students who were enrolled for the third–year course in information systems completed the online questionnaire, yielding a response rate of 22 per cent.

Marks comparison

The assignments of the 111 students who responded to the questionnaire were remarked by the lecturer using the same rubric.

The peer assessment process required each project to be assessed by two students after which the average of the two marks was awarded as the final mark for the project. For further analysis, the assessors were classified as Student 1 and Student 2 for each project. The minimum, maximum, mean and standard deviation of the marks awarded by the first student (PA1), second student (PA2), the average mark of the two student assessments (PA), and the mark awarded by the lecturer (LA) were as shown in Table 1.

Table 1. Marks Comparison

	N	Min	Max	Mean	Std Dev
MA–R					
Peer 1 assessment (PA1)	111	50	100	83.96	12.52
Peer 2 assessment (PA2)	111	50	100	82.88	13.03
Peer assessment AVG (PA)	111	50	100	83.42	11.08
Lecturer assessment (LA)	111	50	100	76.13	13.83
AU–WP					
Peer 1 assessment (PA1)	110	51.25	100	89.25	11.29
Peer 2 assessment (PA2)	110	60	100	89.72	10.76
Peer assessment AVG (PA)	110	56.25	100	89.48	10.69
Lecturer assessment (LA)	110	46.25	100	91.82	11.49

For MA–R the average awarded by the students (PA: 83.42 per cent) was 7.29 per cent higher than the average mark awarded by the lecturer (LA: 76.13 per cent). For AU–WP the average awarded by the students (PA: 89.48 per cent) was 2.34 per cent lower than the average mark awarded by the lecturer (LA: 91.82 per cent).

To determine the reliability of the mark awarded by students, the intraclass correlation coefficient (ICC) and the standard error of the mean (SEM) was calculated.

The ICC determines the reliability of ratings by comparing the variability of different ratings of the same individuals to the total variation across all ratings and all individuals (Shrout and Fleiss 1979):

- A high ICC (close to 1) indicates high similarity between values from the same group.
- A low ICC (ICC close to zero) means that values from the same group are not similar.

(Koo and Li (2016) suggest the following for interpreting ICC:

Below 0.50: poor

Between 0.50 and 0.75: moderate

Between 0.75 and 0.90: good

Above 0.90: excellent

For the standard error of the mean (SEM), the value indicates how far sample means are likely to fall from the population mean using the original measurement units. Larger values correspond to wider distributions.

The results for the two assignments are shown in Table 2.

Table 2. Reliability of peer assessments

	A lect vs stud avg		B lect vs stud1		C lect vs stud2	
	MA–R	**AU–WP**	**MA–R**	**AU–WP**	**MA–R**	**AU–WP**
ICC (agreement)	0.50	0.79	0.37	0.76	0.54	0.77
SEM	8.06	4.88	9.93	5.25	8.48	5.22

Column A shows a comparison of the final mark awarded to the students (the average of the two peer raters) and the mark determined by the lecturer. The ICC of 0.50 for MA–R indicates moderate reliability while the ICC of 0.79 for AU–WP indicates good reliability. The higher SEM of MA–R (8.06) compared to AU–WP (4.88) gives further support for the argument that the peer assessment for AU–WP was more reliable than that of MA–R.

To determine whether the rating of a single student was as reliable as the average of the two raters, the lecturer's mark was fist compared to that awarded by the first group of students (stud1; column B) and then with the marks awarded by the second group of students (stud2; column C). For AU–WP, the reliability for both groups was more than 0.75 which indicates good reliability. For MA–R, even though the ICC for comparison with stud2 was 0.54, which indicates moderate reliability, the ICC for comparison with stud1 was 0.37, which indicates poor reliability. This means that for MA–R, the average rating of two students significantly reduced the risk of unreliable ratings, compared to using the rating of just one student.

One of the reasons for the lower reliability of MA–R compared to AU–WP could be that for MA–R more subjective judgement was required in the marking rubric. The marking rubric for AU–WP consisted of 28 questions with very specific criteria which had to be evaluated, e.g. "Did the student test for obsolete stock?" The marking rubric for MA–R consisted of 7 questions which totalled up to 10 marks. Six questions required very specific criteria to be evaluated but the remaining question required the assessor to give a mark out of 3 for "General Impression". The total difference on the 6 specific criteria questions contributed 1.5 per cent

while the impression mark contributed 5.79 per cent of the total difference of 7.29 per cent (83.42 per cent vs 76.13 per cent).

When criteria on the rubric was very specific requiring little judgement, there was a small difference between the marks awarded by the students compared to that of the lecturer. When more subjective judgement was required, students tended to award higher marks than the lecturer.

Questionnaire results

The following statements were formulated which required students to indicate their agreement with a statement on a 5–point Likert scale where 5 indicates strong agreement and 1 strong disagreement:

"The students that graded my work understood the work I submitted."

"I was satisfied with the mark I received from my peers."

"I believe my peers awarded me a better mark than what the lecturers would have given me."

"The marking grid equipped me to mark the assignments competently."

The following request was posed to gain an understanding of how students experienced the peer assessment process:

"Please provide us with additional comments so that we can understand your experience regarding the peer assessment process better."

The means and standard deviations for each of the Likert–scale questions were calculated and are presented in Table 3.

The answers to the open–ended questions were analysed to identify common themes after which the occurrence of common themes was quantified. Typical answers from students are reported under the qualitative analysis section of this article.

The results of the questionnaire were as follows:

Table 3. Questionnaire results analysis

		Disagree		N	Agree			
		1	**2**	**3**	**4**	**5**	**AVG**	**STD Dev**
The students that graded my work understood the work I submitted.	AU–WP	1%	5%	27%	45%	21%	3.8	0.87
		6%		27%	66%			
	MA–R	1%	6%	30%	44%	19%	3.75	0.86
		6%		30%	63%			
I was satisfied with the mark I received from my peers.	AU–WP	2%	4%	9%	45%	41%	4.19	0.88
		5%		9%	86%			
	MA–R	3%	5%	13%	43%	37%	4.07	0.96
		7%		13%	80%			
I believe my peers awarded me a better mark than what the lecturers would have given me.	AU–WP	7%	25%	25%	24%	18%	3.20	1.22
		33%		25%	42%			
	MA–R	10%	21%	29%	24%	16%	3.15	1.22
		31%		29%	40%			
The marking grid equipped me to mark the assignments competently.	AU–WP	0%	6%	19%	39%	36%	4.05	0.90
		6%		19%	75%			
	MA–R	0%	5%	19%	46%	30%	4.00	0.84
		5%		19%	76%			

For both assignments most students agreed (by selecting 4 or 5) that their fellow students understood their work, that they were satisfied with their marks, and that the marking grid had equipped them to mark the assignments competently.

As far as whether they believed their peers had awarded them a better mark than what the lecturer would have done, 42 per cent and 40 per cent respectively agreed (choosing 4 or 5 on the Likert scale), while 33 per cent and 31 per cent respectively (choosing 1 or 2 on the Likert scale) disagreed, with the rest being neutral (choosing 3 on the Likert scale). Even though the lecturer and students had awarded similar marks for AU–WP, 42 per cent of the students felt that their peers had awarded them a higher mark. Cho, Schunn and Wilson (2006) also found that even though peer ratings were reliable, students' estimates of those ratings were low.

Difference between two assignments

A mixed–model ANOVA test was used to determine whether there was a significant difference between the responses for AU–WP and MA–R respectively. Although the average level of agreement was slightly higher for all the statements for AU–WP, the differences were not statistically significant for any of the statements.

Qualitative analysis

Students were also asked to provide comments on their perceptions regarding the reliability of the peer assessment process. Some of the relevant negative comments were as follows (18 negative remarks regarding the reliability of peer assessment were made):

> "I certainly gave the students I marked the "benefit of the doubt" and I'm sure I was given the same. In this sense, it brought students together, but I do feel our marks may have in general been over–stated. "
>
> " Students are more lenient towards each other."
>
> "Peer assessments would be more beneficial should there be more accountability of the students marking the assignments. More emphasis needs to be placed on the fact that students need to read and understand the assignment they're marking and then give a mark. The memos are also too basic and do not account for all answers."
>
> "The only difficulty is that everyone can interpret certain aspects differently as well as just to ignore the prescribed process."
>
> "… too much room for bias thus negative aspect of the process"

Some of the comments of students who were unhappy with their marks were as follows:

> "I put in a lot of effort whilst doing my work which was heavily under graded by my peers."
>
> "I think the peer assessment should be reviewed by a lecturer as for the working article I had included all the information and the peer that marked my assignment managed to mark me down however I had met all criteria required to achieve the maximum mark"
>
> "I was unsure if my peer had graded me correctly and I asked for a remark. The mark that a got from my peer was 65; after the remark my mark was 90 per cent."

As far as the use of the rubric is concerned, it is evident that the students realised that more judgement was required for MA–R:

> "The working articles were easier to mark and understand as they had set answers but the reports were not fair because the content didn't necessarily mean you met the standards nor did it mean that you answered the given questions to the best standard."
>
> "A lot of students did not take the time to read the whole report and just gave a mark based on the impression of the report. The working articles were easier to assess as the guideline was more strict on specific calculations."
>
> "Only the working articles were straightforward but the reports were not."

Some other comments about the rubrics were noteworthy:

> "Perhaps it would have helped to provide a "model answer" together with the marking grid to make us a little more confident that we were marking fairly."
>
> "Provide students with a clearer picture of how the marking criteria will be structured."
>
> "There is a level of unsurety when you mark others and professional judgement is required. Making the marking grid a little bit more broad allowing us to read on there and look for the point in the work."

Some comments about the process not being anonymous are also insightful:

> "It's great except for the fact that it was not anonymous. The only thing which would give students anxiety is if it was not anonymous."
>
> "… an anonymous peer review process might be more positively perceived by the class."
>
> "… it could be more helpful to remove the names from the documents being assessed, because students tend to mark more leniently if they know the student whose work they are marking."
>
> "Peer review is very subjective I feel because people who know each other in res might be kinder to each other than when they are marking a random student that they didn't even know was in their class."

Where less subjective judgement was required in the marking rubric, the reliability of the peer assessment was good (for AU–WP) compared to reliability being moderate where more subjective judgement was required (for MA–R). Students tended to award higher marks than the lecturer where more subjective judgement was required. While the difference in the average response of students on whether the rubric had equipped them to mark the assignments competently was not statistically significant, it is noteworthy that 36 per cent of students chose a 5 on the Likert scale for AU–WP, compared to 30 per cent for MA–R. The remarks made by students as part of the qualitative analysis supports the finding that students tended to be more lenient where more subjective judgement was required and that they also experienced assessing their peers to be more difficult when more subjective judgement was required of them.

Given that the weighting for AU–WP and MA–R was 17 per cent and 8 per cent respectively towards the predicate mark (which contributes 50 per cent towards the final mark awarded for the course) of the students, the reliability is acceptable in this context. For example, if a student got 10 per cent more for MA–R, the effect on the predicate mark would be 0.8 per cent and 0.4 per cent on the final mark. While 42 per cent and 40 per cent of students believed

that their peers had awarded better marks than what the lecturer would have awarded for AU–WP and MA–R respectively, it was on average only true for MA–R. It must also be noted that a different lecturer could also have given different marks where subjective judgement was required (Charamba and Dlamini–Nxumalo 2022). Reliability of grading by lecturers could also suffer due to a shift in criteria if large numbers of assignments are graded (Cho, Schunn and Wilson 2006). The fact that grading is not an exact science further strengthens the argument that the reliability of the assessment done by peers is adequate and that it can safely be employed in this context.

Some of the lessons learned in the process was that at the time the peer assessment took place, the teaching and learning support system that was in use did not provide for anonymous allocation of peer assessors. Although the random allocation of assessors in such a big group of students probably meant that many students did not know their assessors, there could have been instances where the objectivity of students might have been influenced if they were required to evaluate someone they knew well. While it was not the focus of this study, some evidence of potential biased assessment due to familiarity between the assessor and the assessed was found in the comments made by students. When functionality allows it, we prefer to keep the process anonymous.

The peer assessment process can also be improved by training students to make better subjective judgements. Giving examples of different levels of attainment could possibly improve the reliability of subjective grading.

The number of peer evaluators, especially where more subjective judgement is required, could potentially improve the reliability of the final mark awarded. This matter should however be considered carefully as it will increase the workload of students who might then rush the process.

CONCLUSION

This study was conducted to evaluate the reliability of peer assessment in a third–year information systems course. Two assignments were peer–assessed by students using an online rubric for each assignment. For the one assignment a higher percentage of the marks required subjective judgement compared to the other assignment. The assignments were also assessed by the lecturer and the students' marks were compared with those awarded by the lecturer. While the reliability of both assignments was found to be acceptable, students tended to give higher marks than the lecturer where the rubric required more subjective judgement. A survey among the students revealed that most students agreed that their fellow students understood their work,

that they were satisfied with their marks and that the marking grid had equipped them to mark the assignments competently. Although a significant number of students believed their peers had awarded them higher marks than what the lecturer would have awarded for both assignments, this was only true for the assignment requiring more subjective judgement.

This study provides motivation for lecturers to employ peer assessment for assignments as it is sufficiently reliable, provided that properly designed rubrics are used, and students are adequately guided through the process.

Further research opportunities include determining whether the reliability of peer–assessment is improved if students are given examples of what constitutes different levels of achievement where subjective judgement is required. The effect of anonymity and more evaluators per effort on the reliability of peer assessment is also worthy of further research.

REFERENCES

Andrade, H.G. 2005. Teaching with rubrics. *College Teaching* 53(1):27–30.

Ballantine, J.A. and P. M Larres. 2007. "Final year accounting undergraduates' attitudes to group assessment and the role of learning logs." *Accounting Education* 16(2):163–183.

Barac, K., M. Kirstein and R. Kunz. 2021. "Using peer review to develop professional competencies: an Ubuntu perspective." *Accounting Education* 30(6):551–577.

Bennett, C. 2016. "Assessment rubrics: thinking inside the boxes." *Learning and Teaching* 9(1):50–72.

Berry, F.C., W. Huang and M. Exter. 2023. "Improving Accuracy of Self–and–Peer Assessment in Engineering Technology Capstone." *IEEE Transactions on Education* 66(2):174–187.

Bloxham, S., B. den–Outer, J. Hudson and M. Price. 2016. "Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria." *Assessment and Evaluation in Higher Education* 41(3):466–481.

Boud, D., R. Cohen and J. Sampson. 1999. "Peer Learning and Assessment." *Assessment & Evaluation in Higher Education* 24(4):413–426.

Charamba, E. and N. Dlamini–Nxumalo. 2022. "Same yardstick, different results: efficacy of rubrics in science education assessment." *EUREKA: Social and Humanities* (4):82–90.

Cho, K., C. D. Schunn 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):891–901.

Cockett, A. and C. Jackson. 2018. "The use of assessment rubrics to enhance feedback in higher education: An integrative literature review." *Nurse Education Today* 69:8–13.

Dale–Jones, G., P. Hancock and K. Willey. 2013. "Accounting Students in an Australian University Improve their Writing: But How Did it Happen?" *Accounting Education* 22(6):544–562.

Darvishi, A., H. Khosravi, S. Sadiq and Gašević, D. 2022. "Incorporating AI and learning analytics to build trustworthy peer assessment systems." *British Journal of Educational Technology* (December 2021):844–875.

Fagot, R.F. 1991. "Reliability of Ratings for Multiple Judges: Intraclass Correlation and Metric Scales." *Applied Psychological Measurement* 15(1):1–11.

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.

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.

Hassan, O., A. Fox and G. Hannah. 2014. "Self– and Peer–Assessment: Evidence from the Accounting and Finance Discipline." *Accounting Education* 23(3): 225–243.

Jones, I. and L. Alcock. 2014. "Peer assessment without assessment criteria." *Studies in Higher Education* 39(10):1774–1787.

Karaman, P. 2024. "Effects of using rubrics in self–assessment with instructor feedback on pre–service teachers' academic performance, self–regulated learning and perceptions of self–assessment." *European Journal of Psychology of Education* (June, 29).

Kilgour, P., M Northcote, A. Williams, and A. Kilgour. 2020. "A plan for the co–construction and collaborative use of rubrics for student learning." *Assessment and Evaluation in Higher Education* 45(1):140–153.

Kollar, I. and F. Fischer. 2010. "Peer assessment as collaborative learning: A cognitive perspective." *Learning and Instruction* 20(4):344–348.

Koo, T.K. and M.Y. Li. 2016. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine* 15(2):155–163.

Langan, A.M., C.P. Wheater, E.M. Shaw, B.J. Haines, W.R. Cullen, J.C. Boyle, D. Penney, J.A. Oldekop et al. 2005. "Peer assessment of oral presentations: Effects of student gender, university affiliation and participation in the development of assessment criteria." *Assessment and Evaluation in Higher Education* 30(1):21–34.

Li, H., Y. Xiong, X. Zang, M.L. Kornhaber, Y. Lyu, K.S. Chung and H.K. Suen. 2016. "Peer assessment in the digital age: a meta–analysis comparing peer and teacher ratings." *Assessment and Evaluation in Higher Education* 41(2):245–264.

Lin, G.Y. 2018. "Anonymous versus identified peer assessment via a Facebook–based learning application: Effects on quality of peer feedback, perceived learning, perceived fairness, and attitude toward the system." *Computers and Education* 116:81–92.

Liu, N.F. and D. Carless. 2006. "Peer feedback: The learning element of peer assessment." *Teaching in Higher Education* 11(3):279–290.

Luaces, O., J. Díez and A. Bahamonde. 2018. "A peer assessment method to provide feedback, consistent grading and reduce students' burden in massive teaching settings." *Computers and Education* 126(July):283–295.

Lundstrom, K. and W. Baker. 2009. "To give is better than to receive: The benefits of peer review to the reviewer's own writing." *Journal of Second Language Writing* 18(1):30–43.

Matherly, M. and L. Burney. 2009. "Using peer–reviewed writing in the accounting curriculum: A teaching note." *Issues in Accounting Education* 24(3):393–413.

McDonald, B. and D. Boud. 2003. "The impact of self–assessment on achievement: The effects of self–assessment training on performance in external examinations." *Assessment in Education: Principles, Policy and Practice* 10(2):209–220.

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.

McIsaac, C.M.P. and J.F. Sepe. 1996. "Improving the writing of accounting students: A cooperative venture." *Journal of Accounting Education* 14(4):515–533.

Nawas, A. 2020. "Grading anxiety with self and peer–assessment: A mixed–method study in an Indonesian EFL context." *Issues in Educational Research* 30(1):224–244.

Nicol, D. and D. MacFarlane–Dick. 2006. "Formative assessment and selfregulated learning: A model and seven principles of good feedback practice." *Studies in Higher Education* 31(2):199–218.

Nicol, D., Thomson, A. and C. Breslin. 2014. "Rethinking feedback practices in higher education: a peer review perspective." *Assessment and Evaluation in Higher Education* 39(1):102–122.

O'Donovan, B., M. Price and C. Rust. 2004. "Know what I mean? Enhancing student understanding of assessment standards and criteria." *Teaching in Higher Education* 9(3):325–335.

Orsmond, P., S. Merry and K. Reiling. 1996. "The Importance of Marking Criteria in the Use of Peer Assessment." *Assessment and Evaluation in Higher Education* 21(3):239–250.

Orsmond, P., S. Merry and K. Reiling 2000. "The use of student derived marking criteria in peer and self–assessment." *Assessment and Evaluation in Higher Education* 25(1):23–38.

Ortega–Ruipérez, B. and J.M. Correa–Gorospe. 2024. "Peer assessment to promote self–regulated learning with technology in higher education: systematic review for improving course design." *Frontiers in Education* 9(1):1–11

Panadero, E., A. Jonsson., L. Pinedo and B. Fernández–Castilla. 2023. "Effects of Rubrics on Academic Performance, Self–Regulated Learning, and self–Efficacy: a Meta–analytic Review." *Educational Psychology Review* 35(4).

Petkov, D., O. Petkova, M. D'Onofrio and A.T. Jarmoszko. 2008. "Using Projects Scoring Rubrics to Assess Student Learning in an Information Systems Program." *Journal of Information Systems Education* 19(2):241–251.

Phillips, F. 2016. "The power of giving feedback: Outcomes from implementing an online peer assessment system." *Issues in Accounting Education* 31(1):1–15.

Ramon–Casas, M., N. Nuño, F. Pons and T. Cunillera, T. 2019. "The different impact of a structured peer–assessment task in relation to university undergraduates' initial writing skills." *Assessment and Evaluation in Higher Education* 44(5):653–663.

Riley, T. J., and K.A. Simons. 2013. "Writing in the Accounting Curriculum: A Review of the Literature with Conclusions for Implementation and Future Research." *Issues in Accounting Education* 28(4):823–871.

Saito, H. and T. Fujita. 2004. "Characteristics and user acceptance of peer rating in EFL writing classrooms." *Language Teaching Research* 8(1):31–54.

Shrout, P.E. and J.L. Fleiss. 1979. "Intraclass Correlations: Uses in Assessing Rater Reliability." *Psychological Bulletin* 86(2):420–428.

Son, M. and M. Ha. 2024. "Effects of Descriptive Peer Assessment on the Learning of Middle and High School Students in South Korea." *SAGE Open* 14(2).

Sridharan, B., M.B. Muttakin and Mihret, D.G. 2018. "Students' perceptions of peer assessment effectiveness: an explorative study." *Accounting Education* 27(3):259–285.

Sridharan, B., J. Tai, and D. Boud. 2019. "Does the use of summative peer assessment in collaborative group work inhibit good judgement?" *Higher education* 77:853–870.

Sudhakar, A., J. Tyler and J. Wakefield. 2016. "Enhancing student experience and performance throuh peer–assisted learning." *Issues in Accounting Education* 31(3):321–336.

Sullivan, G.M. 2011. "A Primer on the Validity of Assessment Instruments." *Journal of Graduate Medical Education* 3(2):119–120.

Sung, Y.T., K.E. Chang, T.H. Chang, and W.C. Yu. 2010. "How many heads are better than one? The reliability and validity of teenagers' self– and peer assessments." *Journal of Adolescence* 33(1):135–145.

Taylor, B., F. Kisby, and A. Reedy. 2023. "Rubrics in higher education: an exploration of undergraduate students' understanding and perspectives." *Assessment and Evaluation in Higher Education* 48(4)

To, J., E. Panadero and D. Carless. 2022. "A systematic review of the educational uses and effects of exemplars." *Assessment and Evaluation in Higher Education* 47(8):1167–1182.

Topping, K. 1998. "Peer assessment between students in colleges and universities." *Review of Educational Research* 68(3):249–276.

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.

Westermann, K.D., J.C. Bedard and C.E. Earley. 2015. "Learning the "Craft" of Auditing: A Dynamic View of Auditors' On–the–Job Learning." *Contemporary Accounting Research* 32(3):864–896.

Willey, K. and A. Gardner. 2010. "Investigating the capacity of self and peer assessment activities to engage students and promote learning." *European Journal of Engineering Education* 35(4):429–443.

Yan, D. 2024. "Rubric co–creation to promote quality, interactivity and uptake of peer feedback." *Assessment and Evaluation in Higher Education* 49(4)

Zhang, W., Y. Li and W. Zhang. 2024. "More pain, more gain? Extricating the effect of student involvement in rubric co–construction." *Assessment and Evaluation in Higher Education* 49(4)