 Int J Edu Sci, 9(1): 81-92 (2015)

Exploring the Reliability of Self-Assessment and Peer-Assessment in Oral Presentations in Economics: A Sample of Postgraduate Students at a South African University

Josue Mbonigaba[a*] and Saidou Baba Oumar[b]**

[a]University of KwaZulu-Natal, Durban, KwaZulu-Natal, South Africa
[b]University of Buea, Cameroon

KEYWORDS Education. Research. Evidence. Multiracial. Kwazulu-Natal. Cohort

ABSTRACT This paper explores the reliability of self- and peer-assessment at the University of KwaZulu-Natal, in a context of perceived negative intra-class relationships, using data collected from multiracial cohorts of postgraduate students in economics over the period 2007–2013. The analysis is done with descriptive and inferential methods in which reliability of the marks from these assessments is judged in relation to the lecturer's marks. While peer-assessment marks agree in ranking pattern with the lecturer's marks overall, self-and peer-assessment marks are biased in an undiscernible pattern in each of the racial groups making up the sample. These results imply that caution should be exercised in using these assessments for marks in contexts where there are perceived intra-class negative connections.

INTRODUCTION

Self-and peer-assessment in higher education have been used for two purposes: notably enhancing competency-based learning, and aiding in producing students' marks. Using self- and peer-assessment in marking has been achieved by asking students to provide a mark for their own or peers' work. Even if concerns such as social loafing, free-riders and interaction disabilities may affect the first purpose (Salomon and Globerson 1989), for the second purpose a crucial issue has been the reliability of the resulting mark. The literature documents that self-assessment and peers-assessment marks can be biased, depending on factors such as overconfidence in self-marking (Dunning et al. 2004), and commonly non-academic considerations in peer marking. Even if such assessments have become appealing in contemporary education literature with respect to enhancing competence-based learning, the evidence regarding their reliability in marking has been mixed. An aspect that has been often overlooked in this respect has been the reliability of the marks produced in such assessments in a multiracial class setting in the context of an antagonistic past, where self-assessment and non-blinded peer-assessment marks are likely to be affected by historical race relations. This paper contributes to the evidence by exploring the reliability of such assessments in oral presentations in a multiracial class of postgraduate students in economics at the University of KwaZulu-Natal (UKZN) in South Africa.

The benefits of self- and peer-assessment in higher education cannot be emphasised enough. These two types of assessment have been acclaimed in contemporary higher education because they engage students in active learning (see for instance, Boud and Falchikov 2006; Kirby and Downs 2007, among many other studies). Active learning is part of constructivist learning and teaching practices theories, which the literature describes as engaging students in information processing, reflective skills, problem solving and high-level long-term professional competencies (Brew 1995; Boud and Falchikov 2006; Kirby and Downs 2007; Galbraith et al. 2008; Lew et al. 2010; Spiller 2012; Boud et al. 2013; Boud et al. 2014). Furthermore, there is evidence that these assessments have been enjoyable for students (Stefani 1994: 73), although recently students felt that the process is transferring onto them the onerous responsibility of marking (Cassidy 2008: 508). These assessments are consonant with changing expectations of graduates in the workplace and are believed to

Address for correspondence
Josue Mbonigaba
Box 61133, Bishopsgate, Durban
South Africa
Telephone: +27312602503
Fax: +27312607871
E-mail: mbonigaba@ukzn.ac.za

instil lifelong learning and the ability to work in teams. Practically the two types of assessments are guided by the same principles where students self-assess or assess the work of peers on the basis of preset criteria. This is done through a process by which students make a judgement on the extent of self- or peer-performance (Andrade and Du 2007: 160). The literature also documents that self- and peer-assessments increase efficiency in the use of staff time when such assessments result in the production of marks (Boud and Falchikov 1989: 530; Hanrahan and Isaacs 2001:55). In this regard, however, other literature (Stefani 1994:75; Falchikov 2001) cautions that a good self- and peer-assessment practice can eat up significant teacher's time when preparing and implementing these assessments. Another potential benefit is that self- and peer-assessment were found to be instrumental in allocating a composite group work mark to individual students (Freeman 1995; Rafic and Fullerton 1996; Spatar et al. 2015: 372).

Early reviews of self- and peer-assessment studies tended to support the reliability of these assessments. In a review of studies focusing on how the marks of the lecturers compare to the marks of self-marking, Boud and Falchikov (1989) found that there was an agreement between self-assessment and teachers marks in many studies, but noted that the concept 'agreement' was vague. They further found that strong students tended to underrate themselves with the opposite being true among weak students. Stefani (1994), focusing the analysis across a range of subjects and involving students in setting up assessment criteria, reached similar findings, albeit with no evidence of underrating among strong students or overrating among weak students (p. 72). In subsequent studies the main finding was that in only a minority of studies does the validity of these assessments fail (Topping 1998; Dochy et al. 1999; Falchikov and Goldfinch 2000). In a review by Sadler and Good (2006), 70 percent of the reviewed studies suggested that self-assessment is a valid tool to allocate marks. This validity was further documented in the high school sector (Tseng and Tsai 2007). Some of these studies (Boud and Falchikov 1989; Falchikov and Goldfinch 2000, for instance) noted, however, specific tendencies, notably that good students tended to underrate themselves, mature students tended to be more accurate, and students tended to overrate themselves more generally when the marks were to be recognised and when the process was more academic rather than professional practice. The review studies also noted the inappropriateness and inconsistency in the methodologies applied by reviewed studies.

While one strand of studies emphasised the reliability (validity) of the marks, other studies investigated in greater depth factors affecting the reliability of self- and peer-assessment and how to deal with factors likely to affect this reliability. There are indeed many factors that can bias the outcomes of these assessments, including overconfidence in self-assessment where people tend to exaggerate their knowledge (Dunning et al. 2004). The issues of students being reluctant to be unpleasant to peers in peer assessments, of collusion among students to allocate each other above-average marks in peer assessment, of friendship bonds and enmity among students in peer assessment (Boud 1995: 182; Sadler 2005; Topping 2009: 21) have also been identified as factors compromising the reliability of the process. Furthermore, Bushel (2006) identified bias in performance-ranking where top performers downgrade their closest competitors in peer assessment. Sadler (2009) mentioned non-academic considerations in peer-assessment, while Boud et al. (2013) observed strong students underrating themselves and weak students overrating themselves. Clearly, as has been found in the literature, factors such as differences in contexts, level of courses, performance being evaluated, contingencies associated with outcomes, the training provided to carry out these assessments, as well as other factors, underlie contradictory findings in the literature (Topping 2009: 25; Lawson et al. 2012). Concomitant research to address these factors sought to involve students in the process of self-assessment and peer-assessment with more thoughtfulness (Bloxham and West 2004; Mok et al. 2006) and time inputs (Topping 2009). These discussions hint that the context matters and impacts on the results observed.

Indeed, an often overlooked aspect in this literature is the effect of the design of self- and peer-assessment practices and the effects of intra-class social bonding on the reliability of self- and peer-assessment. Because the direction of the bias with regard to reliability (that is the deviation from or convergence with the marks of the lecturer) cannot be assumed a priori in any

context, the effects need to be explored in peculiar contexts. In South Africa, for example, this exploration is particularly important given the perceived effects of historical race relations (because of apartheid) among different socio-demographic components of a class of students.

Objective

There has not been thus far any paper in South Africa exploring the reliability of self- and peer-assessments in a class with previously antagonistic social groupings and where such assessments are not anonymised. Therefore, the objective of this paper is to explore the reliability of self- and peer-assessment at the University of KwaZulu-Natal, in a context of perceived negative intra-class relationships, using data collected from multiracial cohorts of postgraduate students in economics over the period 2007–2013. The remainder of the paper is structured as follows. The next section presents the methodology of the paper, section 3 presents the results, section 4 discusses the results, section 5 presents the conclusion, section 6 highlights the recommendations, while the last section highlights the limitations of the paper.

MATERIAL AND METHODS

To explore the reliability of peer-assessment and self-assessment the paper compared, in different perspectives, the marks allocated by these two types of assessment to the marks allocated by the lecturer. Such a comparison approach assumed that the marks of the lecturer constituted a best benchmark for reliability (although a doubtful assumption in some contexts) as has been the practice in the literature. The self-assessment, peer-assessment, and lecturer-assessment marks came from records of such marks for oral presentations kept to this effect over the period 2007–2013. These marks were collected as part of a teaching and learning exercise for 238 students from seven successive cohorts of postgraduate students enrolled in the course 'Economics of Health Care' at UKZN, South Africa. In each year, the composite mark from the three types of assessments for oral presentations constituted 5 percent of the final mark in this course. Table 1 shows the distribution of demographic characteristics for the students' sample.

Table 1: Summary of student sample characteristics

Social group	*Male*	*Female*	*Total*
Black	13	42	55 (23%)
Coloured	4	17	
Asian	85	56	141 (59%)
White	10	11	21 (9%)
Total	112 (47%)	126 (53%)	238 100%

For each cohort, a student was asked to choose a topic among the 12 topics to be covered over the 12 weeks of the semester. Since in every cohort the number of students was greater than the number of topics, some students had to work on the same topic, albeit with individual oral presentations. The presenting students were then asked to self-assess their own oral presentation after presenting while the rest of the students were asked to peer-assess the work during the presentation. All three types of assessment were done against preset criteria specified for each of the following intervals of performance: 'fair (20% 39%)', 'good (40% 59%)', 'very good (60%-79%)', and 'excellent (80% - 100%)' (See details in the Appendix).

Two types of analysis were used, namely, descriptive and inferential. In the descriptive analysis, the reliability of self-assessment and peer-assessment marks was explored by examining, in relation to the lecturer's marks, placement of the cohorts in intervals of performance, variability of marks, average marks of different cohorts, and trends of marks from one cohort to another. So, independently, each presenting student was placed in an interval of performance and given a mark by each fellow student (peer-assessment), by him or herself (self-assessment) and by the lecturer (lecturer assessment). Descriptive analysis was conducted at a cohort level and the sample level using average values.

The second analysis consisted of formal tests on data for the whole sample (period 2007–2013) to determine whether or not the marks from the lecturer were indeed statistically different from peer-assessment and self-assessment marks. In this respect, two tests were performed, the t-test and the Wilcoxon signed-rank test. The t-test was used to compare the marks of the lecturer to the marks of the peer-assessment, while the Wilcoxon signed-rank test was used to compare the marks of the lecturer to the marks of the self-assessment. Also, the Wilcoxon signed-rank

test was used to compare the marks from peer-assessment to the marks from self-assessment.

The use of the t-test in comparing two populations requires the populations to be normally distributed and have equal variances. These conditions were fulfilled for peer-assessment marks and lecturer marks, and therefore the t-test was used in their comparison. In contrast, the normal distribution requirement of the t-test was not satisfied for the self-assessment marks, which were skewed to the left. For skewed data, the literature suggests the use of the Wilcoxon signed-rank test and this test was used in all comparisons of marks that involved self-assessment marks.

For the whole sample, the paper tested to see whether or not the lecturer's marks were on the average statistically lower than peer-assessment or self-assessment marks. This one-tailed test was based on the information from descriptive statistics that the lecturer's marks were lower than the marks from self-assessment and peer-assessment. Two tests were conducted. The first (t-test) consisted of testing whether the lecturer's marks were statistically lower than peer assessment marks, and the second (Wilcoxon signed-rank test) consisted of testing whether the lecturer's mark were statistically lower than self-assessment marks. The paper conducted a third test (Wilcoxon signed-rank test) to test whether or not the peer assessment marks were statistically lower than self-assessment marks.

Of crucial interest to this paper was the evidence concerning the reliability of peer- and self-assessment marks in subgroups of students? The class is made up of students from historically antagonistic demographic groups following the history of apartheid, a political system segregating the population according to demographic characteristics. Although apartheid ended in 1994, this paper is motivated by the suspicion that peer assessment might bias groups' marks depending on intra-class groups' connections. For instance, we suspected that groups of students from a given social group might be favoured by groups of students from the same social group. Furthermore, cheating and collusion have been found to be possible in these types of assessment (Boud 1995: 182; Sadler 2005). These collusions among racial groups could in turn trigger a response of self-over makring. The latter could ensue when an assessed anticipate underrating by peers from other racial groups. Therefore, the paper tested to see whether there was bias in the marks assigned to specific groups of students, who were stratified by gender and race groups, to understand whether their marks were higher or lower than the lecturer's marks. The reliability of the marks was questioned if the results in these groups behaved differently from the overall results. The null hypothesis in each testing case was: H_0: the average lecturer's mark is equal to the average marks of a given type of assessment against the alternative: H_a: the average lecturer's mark is lower than the marks of a given type of assessment. All analyses were conducted using STATA and Excel softwares. The results of these analyses are reported next.

RESULTS

In exploring the reliability of self-assessment and peer-assessment, this section starts by examining, in relation to the lecturer's marks, how these two types of assessment place the cohorts' average marks in intervals of performance. Table 2 presents such results.

In Table 2, "x", shown against each type of assessment and under a given interval of performance, means that the marks from that type of assessment place the cohort in that interval of performance. Based on this understanding, one can see that self-assessment marks deviate from the lecturer's marks. Table 2 shows that the self-assessment marks and the lecturer's marks have no cohort of students in common in any interval of performance. In contrast, Table 2 shows that peer-assessment marks and lecturer's marks place some cohorts (2007, 2010, and 2013 cohorts) in the same interval of performance. Furthermore, the results suggest that self-assessment marks place cohorts in higher intervals of performance than lecturer's marks do. Note, however, that when the whole sample is considered, the lecturer's marks and peer-assessment marks place the sample in the same interval of performance. Because self-assessment and peer-assessment marks do not place all cohorts in the same interval of performance as suggested by the lecturer's marks, it can be said that their marks differ more generally from the lecturer's marks.

The results in Table 2 show also the variability of marks around the average marks for each type of assessment. The variability of

Table 2: Comparison of self-assessment and peer-assessment marks to the lecturer's marks

Cohorts	*Type of assessment*	*Average mark*	*Standard deviation*	*Level of performance*			
				80-100	*60-79*	*40-59*	*20-39*
	Per Cohort						
	Self-assessment	81	16	X	-	-	-
2007	Peer-assessment	75	9	-	X	-	-
	Lecturer assessment	68	6	-	X	-	-
	Self-assessment	85	17	X	-	-	-
2008	Peer-assessment	73	10	-	X	-	-
	Lecturer assessment	64	8	-	-	X	-
	Self-assessment	83	16	X	-	-	-
2009	Peer-assessment	76	10	-	X	-	-
	Lecturer assessment	70	7	-	-	X	-
	Self-assessment	86	12	X	-	-	-
2010	Peer-assessment	70	9	X	-	-	-
	Lecturer assessment	62	6	-	X	-	-
	Self-assessment	83	15	X	-	-	-
2011	Peer-assessment	77	9	X	-	-	-
	Lecturer assessment	71	6	-	X	-	-
	Self-assessment	85	16	X	-	-	-
2012	Pee-assessment	74	10	-	X	-	-
	Lecturer assessment	65	7	-	-	X	-
	Self-assessment	84	16	X	-	-	-
2013	Peer-assessment	75	12	X	-	-	-
	Lecturer assessment	67	5	-	X	-	-
	For The Whole Sample						
	Self-assessment	84	17	X	-	-	-
Sample	Peer-assessment	74	10	-	X	-	-
	Lecturer assessment	67	8	-	X	-	-

The sign '-' means no type of assessment marks place a cohort of students in that interval of performance.

marks, measured by standard deviation, reflect how differently members of the cohort perform. Given that this variability across the three types of assessment is analysed on the same cohort or sample, variability of marks reflects variability of these types of assessment in allocating the marks. Keeping this caveat in mind, the results show greater variability in self-assessment marks than in the marks of the other two types of assessment. The evidence of greater variability in self-assessment and peer-assessment marks in relation to the variability of the lecturer's marks suggests the unreliability of the marks from the former two assessments.

Exploring differences in the mean marks on each cohort, results in Table 2 highlight that the average marks of self-assessment and peer-assessment are greater than the average marks of the lecturer's assessment. These results apply also to the whole sample although it is important to note that peer-assessment marks are closer to the lecturer's marks and exhibit a more similar trend with lecturer's marks than self-assessment marks do (see Fig.1).

Could this mean that peer-assessment marks are reliable and self-assessment marks unreliable? Until formal tests are done, this question cannot be answered at this stage. The reliability of peer-assessment marks and self-assessment marks can only be established based on the evidence as to whether or not the average mark of the lecturer is indeed statistically lower than the average mark from the other two types of assessment. This evidence is presented in Table 3.

This evidence from the sample data shows that the average marks from the lecturer are less than the average marks of peer assessment as indicated by a negative t-test statistic (t= -3.228). The p-value of 0.0028 suggests that this difference between the lecturer's marks and peer-assessment marks is statistically significant at a 5 percent level of tolerance, implying that indeed the lecturer's marks are lower than the peer-assessment marks. The same conclusion applies to the other two tests. Specifically, the lecturer's marks are statistically lower than the self-assessment marks on the basis of the Wilcoxon signed-

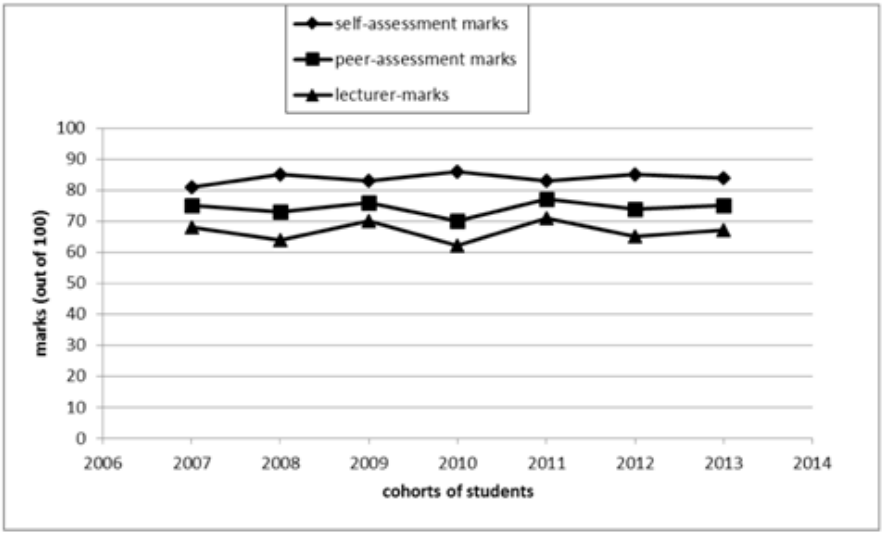

Fig. 1. Trends in marks across three types of assessment

rank test statistic (z = -3.705) with a corresponding p-value of 0.0002, and peer-assessment marks are lower than self- assessment marks on the basis of the Wilcoxon signed-rank test statistic (z = -3.189) with a corresponding p-value of 0.0014.

Table 3: Comparison of peer-assessment marks and self-assessment marks with the lecturer's assessment marks

Lecturer - peer assessment			*Lecturer - self-assessment*			*Peer assessment- self-assessment*		
Left tail test			*Left tail sign test*			*Left tail sign test*		
Hypotheses	*t-stat*	*p-value*	*Hypotheses*	*z-stat*	*p-value*	*Hypotheses*	*z-stat*	*p-value*
Ho: LM-PAM=0			Ho: LM-SAM=0			Ho: PAM-SAM=0	-3.189	0.0014
Ha: LM-PAM<0	-3.228	0.002	Ha: LM-SAM<0	-3.705	0.0002	Ha: PAM-SAM<0		

LM: The lecturer's mark, PAM: peer-assessment mark, SAM: self–assessment mark, N/A: the test is not applicable in this case.
Sources: Estimates obtained with STATA analyses.

Table 4: Comparison of three types of assessment per gender

Assessed gender	*Average mark*							
	Female peer assessors mark	*lecturer's mark*	*comparison*		*Male peer assessors marks*	*lecturer's mark*	*comparison*	
			z	*p-value*			*z*	*p-value*
Female	78.1	73.73	-3.798	0.001	76.43	73.73	-3.125	0.078
Male	74.5	68.06	-3.725	0.0002	75.34	69.80	-3.253	0.000

z: test statistics for the Wilcoxon signed-rank test.
Source: estimates from stats

To test for the reliability of peer assessment, the paper conducted a test to see whether or not the finding for the whole sample was applicable to gender categories of students or whether the trend in the marks of peer-assessment or self-assessment followed the trends in the marks of the lecturer in each gender category. Table 4 shows the results.

Results in Table 4 show that the lecturer's average mark of 73.73 for female students is lower than the female peer assessor's mark of 78.1 for the same gender, and this difference is statistically significant at 5 percent level of tolerance (p-value = 0.0001). Furthermore, the lecturer's average marks of 73.73 for female students is lower than the average marks of the peer male assessors of 76.43 for the same category and this difference is statistically significant at 10 percent significant level (p-value = 0.078). Since these marks are consistent with the overall marks, there is no gender bias in the allocation of marks. This fact is also confirmed by the evidence that trends in marks of female peer assessors and male peer assessors behave similarly across gender. So for instance, it is worth noting that female peer assessors allocate more marks to females (78.1%) than they do to males (74.5%). Likewise, male peer assessors allocate more marks to females (76.43%) than they do to males (75.34%). On the basis of these results, it can be concluded that there is no bias in allocation of marks across gender categories, although the results confirm that peer assessments are not reliable since the lecturer's marks are statistically lower than the peer-assessment marks.

In a sample with multi-social groupings, one would expect social group bias in the allocation of marks. To this end, tests were conducted to compare the average marks allocated by the lecturer to each social grouping and the marks allocated to those groupings from peer assessors from different social groups and the trends in marks of a given grouping. The results of such analysis are presented in Table 5.

The lecturer's marks for different race-based groups were generally lower than the marks allocated to the same groups by peer Asian assessors, except for assessing the White students group. This result reveals that Asian students tended to overestimate the ability of race-based groups, except for the White students group. The results of the marks allocated to different race-based groups by the lecturer are generally

Table 5: Comparison of peer-assessment mrks of social groups and the lecture's mark

Assessed	*Average mark*											
	Asian assessors	*Lecturer*	*p-value*	*Black assessors*	*Lecturer*	*p-value*	*Coloured assessors*	*Lecturer*	*p-value*	*White assessors*	*Lecturer*	*p-value*
Asian	83.63	76.02	0.0000	79.16	74.52	0.0000	47.40	74.68	0.1431	72.29	76.67	0.4213
Black	71.79	70.89	.2430	73.87	69.34	0.0032	69.9	70.14	0.7137	69.76	69.76	1.0000
Coloured	77.3	60.0	0.0048	66.23	69.76	0.254	83	64.28	0.0178	72.25	65.00	0.0947
White	74.42	73.71	0.8190	72.91	66.52	0.0146	71.87	74.40	0.5271	70.50	72.66	0.5282

Coloured: Born from a marriage between a white person and a non-white person (or their descendants). Elsewhere the term "coloured" as used in South Africa, could be defined as "mixed race". The study did not distinguish between the nationalities of students. Source: Authors' based on STATA estimation.

lower than the marks allocated to these groups by Black peer assessors, except for the Coloured group, suggesting that Black students tended to overestimate the ability of other race-based groups, except the Coloured group. The marks allocated by the lecturer to the various race-based groups were higher than the marks allocated to the same groups by Coloured students. This indicates that the Coloured group tended to underestimate the ability of other race-based groups, except themselves. The marks allocated by the lecturer to these race-based groups were higher than the marks allocated to these groups by the White students. This result implies that White students underestimated the intellectual ability of other race-based groups.

Analysing the trends in the marks as one moves across these racial groupings, the results show that Asian assessors rank the groups, from lower to higher, in this order, Black, White, Coloured and Asian, while the lecturers rank them in the order, Coloured, Black, White and Asian. Black assessors rank the groups in this order, Coloured, White, Black, and Asian, while the lecturer's ranking is in the order of White, Black, Coloured, and Asian. Since the ranking of race-based groups in terms of marks is not consistent with the lecturer's ranking, these results indicate that there is a racial bias in the marks assigned by peer-assessment.

Briefly, the evidence presented in the results for race-based groups points not only to inconsistencies between the lecturer's marks for the different groupings and the marks allocated by peer-assessment to these groupings, but that they are also inconsistent with the overall marks trends of the sample. For instance, not all the average lecturer's marks for the different groups were statistically lower than the marks allocated by peer assessors, highlighting that some peer assessors overestimated the ability of students from different groups. This result deviates from the results observed for the whole sample. Although the nature of the bias could not be detected, the results highlight that peer-assessment carries race-based group bias in this sample of students.

DISCUSSION

The use of self-assessments and peer-assessments have been acclaimed in contemporary higher education for two main reasons: their potential to free up marking time for research on the lecturer's side and their pedagogical advantages in nurturing critical thinking and professional development among students. In the most recent literature though, the contention has been that contextual and experimental design factors are likely to affect the marks from these assessments. The implication from this literature is, therefore, that these factors need to be taken into account before the marks arising from these assessments are recognised. One of the lingering questions in this respect is whether or not the reliability of these marks is affected by the contexts in which the assessments are conducted. This paper used a multiracial class in a South African university, where negative relationships between races making up the class is perceived to be one such context because of historical antagonism between races (due to apartheid). The experimental design involved marking a non-anonymised oral presentation where students self-marked themselves after their own presentation and were peer-marked during the presentation. The reliability of the resulting marks was judged with respect to the lecturer's marks.

The main findings of the paper are as follows. The marks allocated from peer- and self-assessment in this sample of students were generally higher than the marks allocated by the lecturer, both at the cohort and sample levels. Both self- and peer-assessment marks overestimated the marks in oral presentations, in relation to the marks allocated by the lecturer. Peer-assessment marks, although also overestimating students' marks in the oral presentations trended more similarly to the lecturer's marks than did the self-assessment (Fig. 1). The reliability of peer assessment was further explored by focusing on whether or not there was a bias in the allocation of marks by assessors in a specific social group, such as gender groups, as they assessed males and females in the sample. The main finding in this respect was that, although marks were overestimated by groups of male and female students, there was no bias in the marks male students allocated to female students and vice versa. This conclusion is arrived at on the basis that trends in these marks were consistent with the trends in marks allocated by the lecturer (Table 4).

Comparing peer-assessment marks of peer assessors of a specific race group recorded for other racial groups in the class (Table 5), the

trend resulted in a mixed picture (overestimation or underestimation). This result suggested additional bias in mark allocations to different race groups although the nature of the bias in terms of whether or not there was a specific group favoured by other groups was not established.

The results observed in this paper do not emerge as a surprise but rather meet the expectations outlined in the paper's initial hypothesis. It was in fact the expectation of the paper that given the historical past between racial groups in the class, bias in peer-assessment would be present because of the possible perpetuation of past racial perceptions in the marking process. In fact, it has been the case in the past that non-white racial groups were looked down on because of apartheid.While the apartheid system favoured the white population, these negative sentiments can be believed to have been prevalent even in non-white social groups. The evidence of observed social connections being stronger within each of the groups in daily life, and in particular, recent observations of tense racial relations in the country as a whole (York 2015), points to the fact that even today these perceptions persist in some people's minds. Bias was expected in settings where peer-assessment was not anonymised in such a setting. While such perceived intra-class relations are expected to bias peer-assessment marks, the question remains as to how these perceived intra-class relationships link to the inflated self-assessment marks observed in the results of this paper. The answer to this lies in the compensatory behaviour of individuals in the face of expected loss. With individual students expecting peers to downgrade them, then in this context of the paper, the natural behaviour is to compensate themselves in their self-marking.

These results and their meaning are not in fact divorced from the rest of the literature. While most of the recent studies focused on pedagogical benefits of the assessments (Spiller 2012; Boud et al. 2013; Boud et al. 2014), there is some evidence related to using these assessments for allocating marks that shows that the reliability of the assessments depends on the context and design of the studies. Indeed, early reviews of the studies on this topic (Topping 1998; Dochy et al. 1999; Falchikov and Goldfinch 2000) hinted at the possibility of the effect of context by showing that good students tended to underrate themselves, mature students tended to be more accurate, and that students tended to overrate themselves more generally when the marks were to be recognised and when the process was more academic rather than in professional practice. Furthermore, successive reviews on the topic, one in 1989 (Boud and Falchikov 1989), another in 1998 (Topping 1998) and yet another more recently by Sadler and Good (2006) constitute an evidence of an unresolved topic. Collusion and peer competition for top place among students noted in recent studies (Sadler 2005; Topping 2009: 21) have been some of contextual factors affecting the reliability of these marks. Other literature inferred that inflated marks in self-assessment are a result of the natural behaviour of human beings who commonly exaggerate their ability and knowledge (Dunning et al. 2004). Specifically in this paper, this exaggeration took place in response to an anticipated downgrading by peers. Presumably, these contextual influences on the reliability of such assessments might underlie the recent focus on factors likely to influence bias, including more time inputs by students and lecturers (Topping 2009) and more thoughts on the process itself (Mok et al. 2006). On a positive note, it is worth noting the recent practice in education that relies on self- and peer-assessment to distribute a composite group mark among individual students making up the groups (Spatar et al. 2015) without referring to contexts. Whether or not previous studies contextualise the reliability of these assessments, no paper analysed the question in a multiracial class with an antagonistic past. The only peculiarity of the results of this paper and its contribution to the literature lies in the fact that it presents intra-class racial self- and peer-marking, showing that even if overall peer-assessment has the same pattern as the teacher's marks as in previous evidence, this 'agreement' between the two marks could be hiding serious bias in intra-class peer assessments.

These findings have important implications with respect to using peer-assessment and self-assessment in allocating marks to students' work. Specifically, in light of the findings, using these types of assessment in allocating marks to students means promoting inaccuracy in measuring students' oral presentations' marks, which may impact negatively on educational outcomes if these marks account for a significant percent-

age of the course. Still, in this experimental paper where these marks constituted an insignificant portion of the course percentage, the evidence shows that these marks were biased. Therefore, while they might be good educational tools, these types of assessment should not be used in allocating marks, particularly in multi-racial classes with possible inter-racial negative relationships.

CONCLUSION

Self-assessment marks are not reliable in oral presentations in classes where there are social groupings with perceived negative relationships and this is also the case for peer-assessment marks. While the bias is present in the overall marks arising from each of the assessments, this bias in peer-assessments is enhanced in intra-class racial groups with perceived negative relationships.

RECOMMENDATIONS

The results of this paper suggest two recommendations. Despite the evidence for the validity of marks from peer-assessment elsewhere, and more recommendations tending towards the use of self- and peer-assessment in producing students' marks, this paper recommends that such use be restricted or used with caution in oral presentations in classes where intra-racial grouping's relations are perceived negatively. The paper recommends also more studies to establish the evidence in this area given the limitation of the present paper with respect to generalisation.

LIMITATIONS OF THE PAPER

The experimental design of the paper, the racial composition of the class that was the subject of this this paper could have exerted a significant effect on the results observed, as noted earlier. Due to profound transformation at UKZN since 2004, the composition of a postgraduate class at UKZN is most likely to be different from a typical class composition in terms of races at many other institutions in the country. Therefore, these results cannot be generalizable to other classes of postgraduate students in economics in South Africa.

ACKNOWLEDGEMENTS

The authors would like to thank in advance the anonymous reviewers for their valuable comments and Carol Brammage for English editing.

REFERENCES

Andrade H, Du Y 2007. Student responses to criteria-referenced self-assessment. *Assessment and Evaluation in Higher Education*, 32(2): 159–181.

Bloxham S, West A 2004. Understanding the rules of the game: Marking peer assessment as a medium for developing students' conceptions of assessment. *Assessment and Evaluation in Higher Education*, 29(6): 721-733.

Boud D 1995. *Enhancing Learning through Self-assessment*. London: Kogan Page.

Boud D, Cohen R, Sampson J 2014. *Peer Learning in Higher Education*. London: Kogan Page Limited.

Boud D, Falchikov N 1989. Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education*, 18(5): 529-549.

Boud D, Falchikov N 2006. Aligning assessment with long-term learning. *Assessment and Evaluation in Higher Education*, 31(4): 399-413.

Boud D, Lawson R, Thompson DG 2013. Does student engagement in self-assessment calibrate their judgment over time? *Assessment and Evaluation in Higher Education*, 38(8): 941-956.

Brew A 1995. What is the scope of self-assessment? In: D Boud (Ed.): *Enhancing Learning through Self-assessment*. London: Kogan Page, pp. 48-63.

Bushell G 2006. Moderation of peer assessment in group projects. *Assessment and Evaluation in Higher Education*, 31(1): 91-108.

Cassidy S 2008. Developing employability skills: Peer assessment in higher education. *Education and Training*, 48(7): 508-517.

Dochy F, Segers M, Sluijsmans DMA 1999. The use of self-, peer and co-assessment in higher education: A review. *Studies in Educational Evaluation*, 24(3): 331-350.

Dunning D, Heath C, Suls JM 2004. Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 5(3): 69-10.

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

Falchikov N 2001. *Learning Together: Peer Tutoring in Higher Education*. London: Routledge Falmer.

Freeman M 1995. Peer assessment by groups of group work. *Assessment and Evaluation in Higher Education*, 20: 289–300.

Galbraith RM, Hawkins RE, Holmboe ES 2008. Making self-assessment more effective. *Journal of Continuing Education in the Health Professions*, 28(1): 20-24.

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

Kirby NF, Downs CT 2007. Self-assessment and the disadvantaged student: Potential for encouraging self-regulated Learning? *Assessment and Evaluation in Higher Education*, 32(4): 475-494.

Lawson RT, Taylor DG, Thompson L, Simpson M, Freeman L et al. 2012. Engaging with graduate attributes through encouraging accurate student self-assessment. *Asian Social Science*, 8(4): 3–12.

Lew MDN, Alwis WAM, Schmidt HG 2010. Accuracy of students' self-assessment and their beliefs about its utility. *Assessment and Evaluation in Higher Education*, 35(2): 135-156.

Mok MMM, Lung CL, Cheng DPW Cheung RHP and Ng ML 2006. Self assessment in higher education: Experience in using a meta-cognitive approach in five case studies. *Assessment and Evaluation in Higher Education*, 31(4): 415-433.

Rafic Y, Fullerton H 1996. Peer assessment of group projects in civil engineering. *Assessment and Evaluation in Higher Education*, 21: 69-81.

Sadler DR 2005. Interpretations of criteria-based assessment and grading in higher education. *Assessment and Evaluation in Higher Education*, 30: 175-194.

Sadler DR, Good E 2006. The impact of self and peer-grading on student learning. *Educational Assessment*, 11: 1-31.

Sadler DR 2009. Indeterminacy in the use of pre-set criteria for assessment and grading in higher education. *Assessment and Evaluation in Higher Education*, 34: 159–179.

Salomon G, Globerson T 1989. When teams do not function the way they ought to. *International Journal of Educational Research*, 13: 89–99.

Spatar C, Penna N, Mills H, Kutija V, Cooke M 2015. A robust approach for mapping group marks to individual marks using peer assessment. *Assessment and Evaluation in Higher Education*, 40(3): 371-389.

Spiller D 2012. Assessment Matters: Self-Assessment and Peer Assessment. Teaching Development| Wâhanga Whakapakari Ako. From <http://www.waikato.ac.nz/tdu/pdf/booklets/9_SelfPeer Assessment.pdf> (Retrieved on 10 December 2013).

Stefani LAJ 1994. Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education*, 19(1): 69-75.

Topping KJ 1998. Peer assessment between students in colleges and universities. *Review of Educational Research*, 68: 249–276.

Topping JK 2009. Peer assessment. *Theory into Practice*, 48(1): 20-27.

Tseng CC, Tsai CC 2007. On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers and Education*, 49: 1161-1171.

York G 2015. Analysis: Two Decades After Apartheid Ended, Racial Tensions Rattling South Africa, The Global and Mail, January 22, 2015 p.3. From <http://www.theglobeandmail.com /news/world/two-decades-after-apartheid-ended-racial-tensions-rattling-south-africa/article22571118/> (Retrieved on 15 April 2015).

APPENDIX

Criteria for assessing presentation

Excellent (80-100%	*Very good 60-79%*	*Good 40-59%*	*Fair 20-39%*
- Discuss a theory relevant and directly related to the topic - Discuss relevant application of the theory with a focus on few studies (3-4) discussed in depth - Eexcellent coherence and logic in the discussion (intro-to conclusion) - Many own critical reflections and opinion about the validity of the theory, relation between results of studies (consistent among themselves and with the theory) - Confident and enthusiastic attitude about your topic and clarity of language	- Discuss theory relevant and directly related to the topic - Discuss relevant application of the theory with a focus on few –studies(2-3) discussed in depth - Some coherence and logic in the discussion (intro-to conclusion),presentation clear - some critical reflection - Limited confidence, enthusiasm and clarity	- Discuss theory relevant and directly related to the topic - Discuss relevant application of the theory with a focus on few -studies discussed in (1-2) depth - Some coherence and logic in the discussion (intro-to conclusion, presentation not very clear - Very limited or no critical reflection - Very limited or no confidence, enthusiasm and clarity	- Discuss theory irrelevantly or in a vague way in relation to the topic in question - Limited discussion with respect to the application of the theory - Inadequate or no coherence or logic in the discussion (intro-to conclusion), presentation unclear - No critical reflection - Very limited or no confidence, enthusiasm and clarity