Accepted version of Sridharan, B., Tai, J. and Boud, D. (published online 25 August 2018). Does the use of summative peer assessment in collaborative group assessment inhibit good judgement? *Higher Education,* DOI: 10.1007/s10734-018-0305-7

Does the use of summative peer assessment in collaborative group work inhibit good judgement?

Bhavani Sridharan, Deakin Business School, Deakin University, Geelong, Australia, +61 3 92517410, bhavani.sridharan@deakin.edu.au, ORCID - 0000-0002-2217-949X

Joanna Tai, Centre for Research in Assessment and Digital Learning, Deakin University, Geelong, Australia, +61 3 924 43780, joanna.tai@deakin.edu.au, ORCID - 0000-0002-8984-2671

David Boud, Centre for Research in Assessment and Digital Learning, Deakin University, Geelong, Australia; Work and Learning Research Centre, Middlesex University, London, UK; Professor Emeritus, Faculty of Arts and social sciences, University of Technology, Sydney, Australia; +61 3 924 68038 / 0417 258650, david.boud@deakin.edu.au, ORCID - 0000-0002-6883-2722

The accuracy and consistency of peer marking, particularly when students have the power to reward (or penalise) during formative and summative assessment regimes, is largely unknown. The objective of this study is to evaluate students' ability and behaviour in marking their peers' teamwork performance in a collaborative group assessment context both when the mark is counted and not counted towards their final grade. Formative and summative assessment data were obtained from 98 participants in anonymous self and peer assessment of team members' contributions to a group assessment in business courses. The findings indicate that students are capable of accurately and consistently judging their peers' performance to a large extent, especially in the formative evaluation of the process component of group work. However, the findings suggest significant peer grading bias when peer marks contribute to final grades. Overall, findings suggest students are reluctant to honestly assess their peers when they realise their actions can penalise non-contributing students. This raises questions about the appropriateness of using peer marks for summative assessment purposes. To overcome the problems identified, this paper proposes a number of measures to guide educators in effectively embedding summative peer assessment in a group assessment context.

Keywords: group assessment; formative and summative assessment; consistency and accuracy; assessment bias

Introduction

Preparing students for the world of future work, i.e. promoting employability, is now an established focus of higher education. In many situations, this includes ensuring students learn how to work in groups. The Australian Graduate outlook survey (GCA 2015), identifies teamwork as the third highest ranked skill that recruiters are seeking from university graduates. National (*Tertiary Education Quality and Standards Agency (TEQSA))* and international accreditation agencies (eg. The Association to Advance Collegiate Schools of Business (AACSB)) and universities have mandated the requirement to teach, assess and evidence team working skills. Many universities have embraced team working skills as one of the key graduate learning outcomes in policy documents (eg. Deakin 2014). Courses have therefore instituted learning tasks and activities which foster these abilities. The intention to pursue such skills though has to be balanced against concerns from students that they can be assessed fairly.

There are dilemmas in asking teachers to assess students on their teamwork abilities. Firstly, teachers are not part of the group, and therefore are unable to verify individual students' contribution to the team. This has been partially resolved by asking students to grade each other, though many teachers shy away from allocating much significance to this grade. Secondly, and perhaps more importantly, there will be no university tutor to assess performance within future work environments, where much work is undertaken in teams. The implication of the employability agenda is that students must be able to judge their own work, and the work of others, including in teamwork situations – and this capacity requires development.

Asking students to assess each other in group work tasks aligns with feedback principles (Nicol and Macfarlane-Dick 2006). It provides assessment on observed work (rather than assumptions on the quality of teamwork based on the product of the team), and provides students with an opportunity to develop their own capacity to make judgements (i.e. their evaluative judgement) (Tai et al. 2016). It also enables students to receive multiple sources of feedback on their work, potentially at multiple time points, affording them opportunities to improve their contributions. The key focus of this study is peer assessment of individual contribution to and performance in a group work assignment, rather than peer assessment of the academic standard of an assignment. The former aims to develop students' soft skills and evaluative judgement while the latter aims to regulate and assure academic standards (Bloxham et al. 2015).

For this to be effective, students must first be able to make relatively accurate judgements, so that the assessment is useful and fair. Discussion, moderation, and providing guides or rubrics may assist with consistency and accuracy, as it has for university tutors/markers (Bloxham et al. 2016). Though much has been said of the inaccuracy of peer marking in comparison to teacher marks (Speyer et al. 2011; Falchikov and Goldfinch 2000), there have been few efforts to improve the value of peer assessment in summative contexts (Steverding, Tyler, and Sexton 2016). The studies that do exist have focussed on essays and qualitative feedback generated (To and Carless 2016; Nicol, Thomson, and Breslin 2014), rather than teamwork, or the use of rubrics.

This paper reports on the ability of students to consistently assess each other on team work skills using a rating scale rubric. We outline the relevant literature on group work assessment, peer and self-assessment. We detail the context in which this study was undertaken, including the affordances of a particular self and peer assessment tool, SPARK^{PLUS}. Results of students' ability and behaviour in peer marking are presented, and

discussed in light of employability, and the need to develop students' evaluative judgement. Strategies for improving group work assessment are proposed.

Background

Group work

Many activities can be considered group work, such as discussions or problem based learning activities. We specifically refer here to tasks where students work together to produce a final piece of work which is assessed, that is, a group assessment. Student and educator perception of such activities vary and it is frequently seen as problematic (Sridharan, Muttakin, and Mihret 2018). Nevertheless, group work and assessments are an authentic activity within many disciplines/fields, mirroring requirements for future work practices. Requirements for effective group work assessment include ensuring that work is observed prior to being assessed, and the fairness, division, and/or allocation of marks. Significant work has been undertaken on algorithms and options for adjusting scores to increase fairness of peer marking through calculations, which seek to enhance accuracy and consistency in peer marking (Sung-Seok 2014; Cheng and Warren 2000). However, focussing on scoring alone may lead to neglect of students' learning potential in group work tasks. Opportunities for formative assessment and feedback are important to engender learning; group work is likely no different from other complex skills or capabilities in requiring practice and refinement.

Formative and summative peer assessment

Peer assessment has been used in a formative sense, providing peers with timely information on their ongoing performance. Depending on the context and activity, the benefits of formative peer assessment include better engagement of students in the learning process (Willey and Gardner 2010); improved student motivation, autonomy and control over learning (Planas Lladó et al. 2014; Wen and Tsai 2006); and critical thinking skills (Topping 2009; Nicol and Macfarlane-Dick 2006).

In contrast, summative peer assessment occurs where students assess each other on the basis of their observed performance by the conclusion of the task or assignment which requires teamwork skills to be employed (Sambell, McDowell, and Montgomery 2013). Depending on its purpose, the assessment may comprise a mark, comments, or both, however it usually involves observing, reading, or interacting with peers in completion of the groupwork assessment task. Use of summative peer assessment may promote 'fairness' in marking (Fellenz 2006); reducing social loafing (Sridharan, Muttakin, and Mihret 2018); and student empowerment and development of professional and lifelong skills (Planas Lladó et al. 2014).

Reliability and accuracy concerns

Despite its documented benefits, the use of self and peer assessment for summative purposes is seen to be problematic due to validity and reliability concerns (Topping 2009; Yao-Ting et al. 2010). One of the main concern is students' ability to reliably and accurately assess their peers' work, which is contingent upon a number of antecedent factors. These include students' understanding of quality, standards and expectations (O'Donovan, Price, and Rust 2004) and the capacity for evaluative judgement (Nicol, Thomson, and Breslin 2014). We note, however, that these problems which plague peer assessment also exist more generally regarding rater judgement (Bloxham et al. 2016). The key difference might lie in concerns regarding fairness, honesty and impartiality (Willey and Gardner 2009). While there are usually few incentives for educators to mismark students, social obligations (eg. peer pressure) amongst students may lead to additional distortions of marks, and

so conducting a peer assessment in a way that reduces these obligations may improve its acceptability, for both students and educators. Vickerman (2009) suggests anonymous marking to enhance accuracy, as students can be assured they will not be identified when indicating peers' underperformance.

Self-assessment

In a situation where peer assessment is used for learning, rather than grade allocation, it is unlikely to function well without some element of informed self-assessment. This is also the case with other forms of assessment, but is particularly important in the event that peer assessments provide information which the student must work through before coming to a conclusion on their own performance.

Self-assessment for summative purposes is a contentious topic in higher education. Concerns arise from the tension between its use as part of a pedagogical strategy, and the potential for self-scoring to contribute to grades, and therefore influence certification outcomes. Furthermore, there are ongoing concerns regarding the inaccuracy of self-assessment for summative purposes. The general consensus, however, is that learning and improving skills in the area which is self-assessed, will result in more accurate self-assessment (Boud and Falchikov 1989). In the context of group work, this suggests that students need practice in group work assessment to be able to accurately assess themselves better, and speaks to the argument that participation is necessary for improving evaluative judgement (Tai et al. 2016). Therefore, students need to participate in group work and have opportunities for formative assessment (of themselves, and by others) prior to a final summative moment.

Strategies

To deal with the challenges of peer assessment, a number of strategies have been postulated. For peer assessment alone, suggestions have included frequent exposure to peer feedback and assessment (Nicol, Thomson, and Breslin 2014; Brutus, Donia, and Ronen 2013; Sadler 2010); explicit specification of criteria, benchmarking exercises (Willey and Gardner 2010) and training and calibration of peer assessment (Loughry, Ohland, and Woehr 2014). The combination of self and peer assessment scores is not a new idea either: the concept has existed though terms have changed (Boud, Cohen, and Sampson 1999). All of these strategies may help students to better understand how they perceive their performance in comparison to peers.

Self and peer assessment technology

A range of programs have been developed to ameliorate these logistical issues, such as SPARK^{PLUS} (http://sparkplus.com.au). SPARK^{PLUS} is an online self and peer evaluation system to provide an opportunity for students to rate themselves and their teammates and provide feedback on how their respective teammates can improve their capability to work in a team environment. This requies developing and embedding a list of criteria measuring teamwork into the system. This program assembles scoring information from team members on each individual's performance, and automatically calculates two factors based on both self and peer assessments (Figure 1). These factors facilitate the moderating of peer assessment scores.

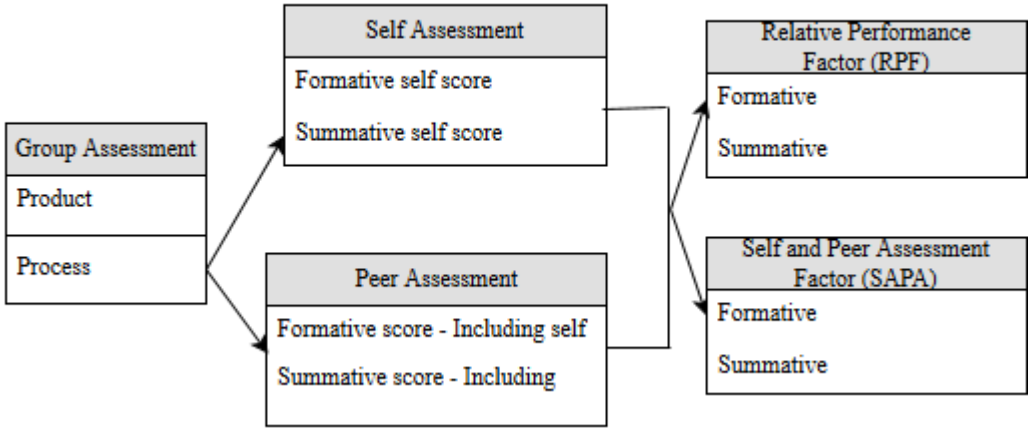

Fig. 1 Group Assessment using SPARK[PLUS]

The first factor, (called RPF) is an indicator of relative contribution to the group, as assessed by how highly an individual in a group is rated compared to the average rating of all group members. A lower number indicates that a student contributed less to the group than other group members. This factor could then be used to multiply the overall team mark to allocate individual marks based on their relative contributions to the team.

The second factor (SAPA) compares the student's own rating of their contribution, to how their team members rated their contribution. A higher number here indicates the student believes they contributed more than their team thinks they did, while a lower number indicates the student thought they contributed less than their peers perceived they did.

These two factors, when presented back to students, can act as a form of feedback on a student's relative contribution to the group work, and the accuracy of a students' self-assessment on their contribution to the group work, relative to other team members' perceptions. By using the information from SPARK[PLUS] in a formative manner, students may be able to better develop their teamwork skills and evaluative judgement around their ability to work in a team.

While tools such as SPARK[PLUS] may reduce logistical and mark calculation issues, there remains a more foundational challenge. Both self and peer assessment are vital to improve students' skills in this group work, and so students must be relatively accurate assessors for this type of assessment both in formative and summative situations. Students must, therefore, have opportunities to develop both their self and peer assessment and group work skills, prior to those skills being used and assessed in a summative situation.

Research Question, Conceptual Model and Hypotheses

This study was driven by the following research question:

> *Can students assess their peers accurately, consistently and without assessment bias irrespective of whether the score is counted or not towards final assessment in a collaborative group assessment context?*

Accuracy refers to validity of peer assessment scores (whether marks given to an individual student by their peers are proportional to their actual/perceived contribution). Consistency refers to reliability in peer marking (whether the peer marking behaviour remains the same for both formative and summative assessment

when graded by the same peer members). Assessment bias refers to the influence or presence of any systematic bias in peer assessment behaviour.

Based on the literature reviewed, we developed three hypotheses (Figure 2) which could be investigated through the use of SPARK^{PLUS} software for peer assessment:

- H1a and H1b: Different levels of actual contribution will lead to varying performance scores (average peer mark) during both (a) summative and (b) formative peer assessment regimes.
- H2a and H2b: Different levels of self-perceived contribution (in comparison to peers) will lead to varying performance scores during both (a) summative and (b) formative peer assessment regimes.
- H3a and H3b: Interdependency between actual contribution and self-perceived contribution levels will influence the performance score during both (a) summative and (b) formative peer assessment regimes.

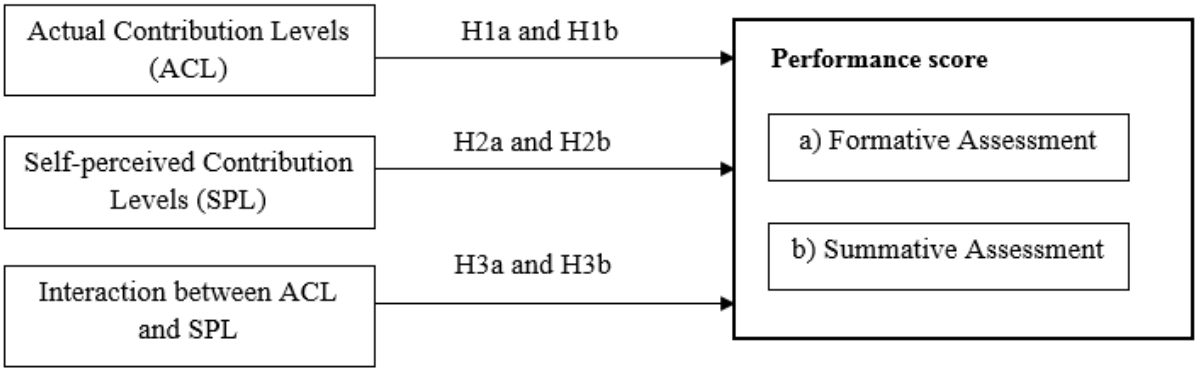

Fig.2 Conceptual Model and Hypotheses

Research Design and Procedures

Participants

The participants for this study were students enrolled in a capstone information systems unit requiring them to apply their acquired knowledge and skills to a real-life project. In particular, this involved completion of industry-based information systems projects using their knowledge and skills in project management, information systems analysis, design and development skills, and professional skills in interacting with project teams. Retrospective data were collected from two consecutive offerings of the unit to undergraduate students (2014 and 2015) and to one offering to postgraduate students (2015) with the same underlying course content in all three. Seventy-one undergraduate students and twenty-seven postgraduate students participated in the peer assessment process, a total of 98 students. Students were provided with resources on developing teamwork skills and an online discussion forum was provided for each team. Although all students had been exposed to a collaborative group work environment prior to this, this was their first occasion of assessment on teamwork.

Assessment Design

The overall allocation of marks for the both undergraduate and postgraduate units were: (a) 80% from a group work product mark and (b) 20% from an individual work mark. This 20% is assessed via a reflective essay based on project diary and team reflection for undergraduate students and project presentation for the postgraduate students.

The two group work components that students were assessed on were: group product and teamwork process. The group product required students to deliver four key components: draft project proposal; system development plan; system design and development; and a final report including user and systems manual and recommendations. These were graded by teaching staff at each stage. The teamwork process was assessed both formatively and summatively by the team members. The formative teamwork assessment process was based on three broad criteria: contribution to ideas; contribution to tasks; and collaborative skills. Students rated self and peers' on the five-point sliding scale option relating to these three assessment criteria. The five scale descriptors were never, rarely, sometimes, often and almost always. The assessment process, access to the SPARK^{PLUS} system, and how interpret the results upon publication were communicated to students before the commencement of the assessment task. The marks allocated for this were primarily aimed at helping students to positively change their teamwork behaviour and these were not counted towards the final grade.

The final summative peer assessment was a holistic assessment by each team member of their own and others' contribution to the group. The final group work product mark was adjusted using the summative peer score resulting in an individual product mark. However, there was a slight variation between how this peer mark was used to adjust individual group work product marks. In 2014, individual marks for the group work component were adjusted based on their individual contribution to teamwork and group work grade weight (80%) (see Figure 3). Here, the individual adjusted product score is obtained by multiplying groupwork product mark by individual peer score (%).

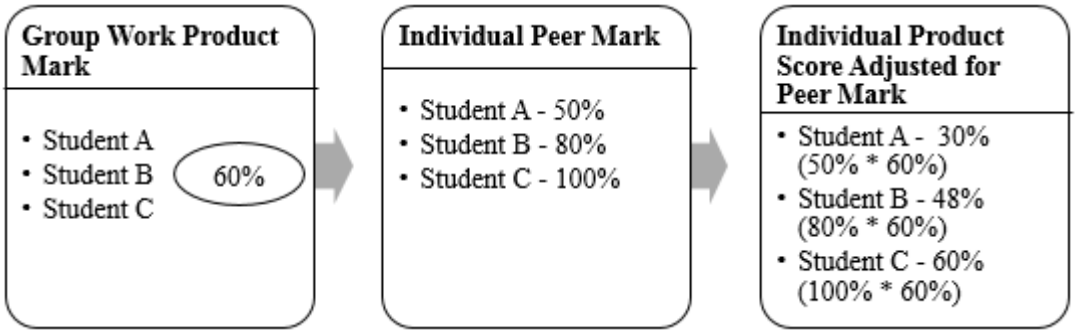

Fig. 3 Example calculations used for adjusting group work product mark

In 2015, the individual mark was adjusted (increased or decreased) at the unit chair's discretion based on multiple sources of information including average peer mark, confidential communication/complaints about team members and observation on their contribution to the discussion forum. This was due to the change of unit chair in 2015. It was believed that postgraduate business students, particularly those already working in the industry, find self-refection a superficial exercise and they had previously expressed dissatisfaction in undertaking such a task. Since this study's focus is on peer assessment of process (teamwork), these adjustments had no bearing on the findings and excluded from statistical analysis.

The sequence of activities in the peer assessment process in this study took place over 11 weeks (Figure 4). Students were allocated randomly into groups by the unit chair with the target of five to seven members for undergraduate units and three to five for postgraduate units. The final range of group size after movement between groups and withdrawals was four to five for undergraduates and three to five for post graduates. Each group was allocated a separate project on which they worked collaboratively. Assessment

specifications and relevant resources for completion of the group work product and teamwork process were provided to students through the learning management system. After students worked as a group and completed their first two tasks, students were given access to the peer assessment system (secure online SPARK^{PLUS} tool) to complete the formative peer assessment cycle after week 4.

The peer assessment process enabled students to rate each other anonymously so as to minimise any inhibition or threat. Release of the formative self and peer assessment results occurred soon after the completion of peer marking to allow students to act on the feedback in the form of peer mark and qualitative peer feedback. Once the final product was delivered, students again completed the final peer marking using the same system and the results and the qualitative peer feedback were published soon after the completion of the summative peer scoring process. Following this, the final assessment task was submitted.

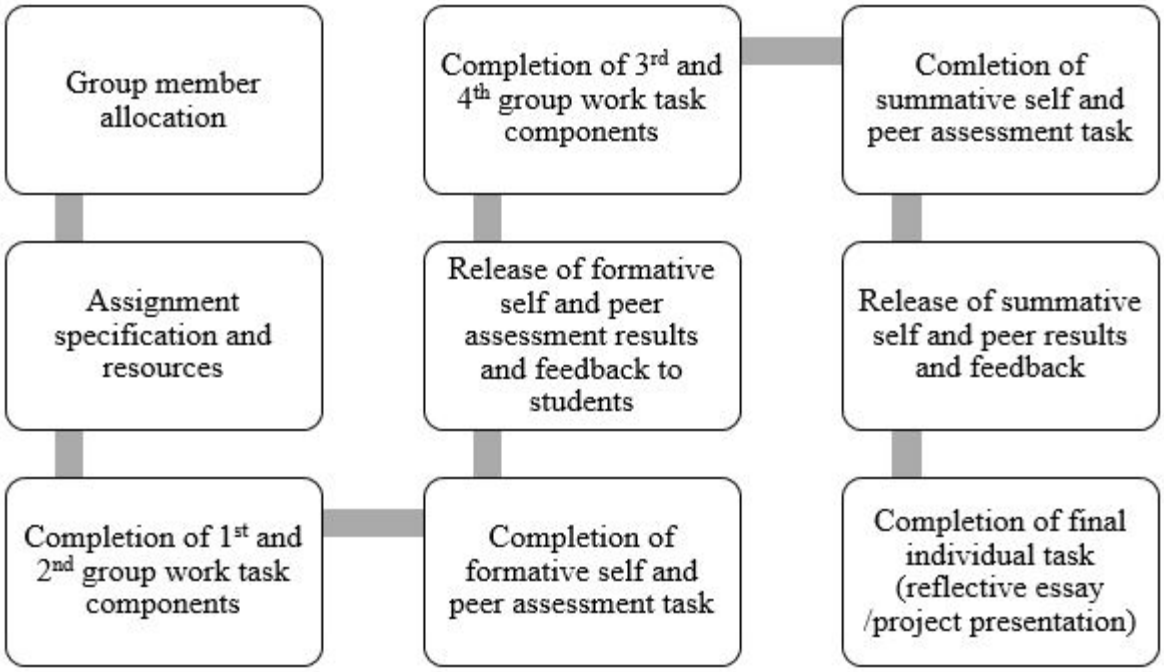

Fig. 4 Sequence of learning activities demonstrating positioning of formative and summative self and peer assessment process

Data Measurement and Analysis

To evaluate accuracy, consistency and bias, data was gathered from formative and summative assessment regimes in the form of self and peer assessment scores, relative performance factor (RPF) and self and peer assessment factor (SAPA) from the online SPARK^{PLUS} tool. Example calculations of these two factors are shown in Figure 5.

$$\text{RPF for Student A} = \sqrt{\frac{\text{Total Mark for Student A}}{\text{Average Group Mark}}}$$

$$\text{SAPA for Student A} = \sqrt{\frac{\text{Self Assessed Mark for Student A}}{\text{Average Peer Mark for Student A}}}$$

Fig.5 Example RPF and SAPA Calculations (Wiley and Garner, 2009)

Table 1 summarises the variables used in this study. The dependent variable is peer assessment performance score with a higher score indicating superior performance during (a) *formative* and (b) *summative regimes*. These scores are calculated as average scores for multiple criteria and from multiple peers. It excludes the self-assessed score so as to remove the effects of inflated self-assessment score bias (Lejk and Wyvill 2001).

The RPF factor and SAPA factors are used to derive the two independent categorical variables used in this study; namely, actual contribution levels relative to overall team contribution (will be referred to as ACL) and self-perceived contribution levels relative to overall team-perception of individual contribution (will be referred to as SPL). The ACL is derived from relative performance factor (RPF) scores for each student. This measures the level of individual student contribution in comparison to the entire team. The ACL's are classified into under-contributor (i.e. individual contributed less than their team), equal-contributor (i.e. individual contributed the same as their team) and over-contributor (i.e. individual contributed more than their team).

The SPL is a derived variable based on self and peer assessment (SAPA) factor scores for each student. This measures the level of self-perceived contribution in comparison to the entire team's perceived individual contributions. The SPL's are classified into overinflated (i.e. individuals rated their contribution more than they were rated by peers), accurate or modest (i.e. individuals rated their contribution similar to or less than they were rated by peers) (Table 1). The rationale for coupling accurate and modest classification is to overcome the limitation of small sample size and to enable further analysis using powerful statistical techniques.

Table 1. Variables and classification levels

Variable	Classification levels
Performance score	Individual student's performance score is the average of peer mark from multiple markers for multiple criteria excluding self-assessment score.
ACL	Individual student's ACL is classified into: under-contributor - RPF values <0.98; equal-contributor - RPF values between 0.98 and 1.02; over-contributor - RPF values > 1.02.
SPL	Individual student's SPL is classified into: overinflated - SAPA values > 1.02; accurate or modest - SAPA values ≤ 1.02.

Prior to conducting the analysis, the data were tested for various assumptions such as non-violation of normal distribution and homogeneity of variances. Skewness and kurtosis values were used to evaluate the non-

violation of normality assumption using the threshold values of ±3 (Gravetter and Wallnau 2005; Holmes-Smith, Cunningham, and Coote 2006). Levene's homogeneity of variance tests was conducted in the first instance. If this assumption was violated, more stringent threshold alpha value of 0.001 along with a Bonferroni adjustments to alpha level was used to reduce the chances of obtaining false positive results (Tabachnick, Fidell, and Osterlind 2001). Following this, a two-way analysis of variance (ANOVA) was conducted for data gathered from both *formative* and s*ummative regimes*. An alpha level of 0.01 or 0.003 (adjusted alpha value based on Bonferroni correction) was used for further analysis. The effect size ($\eta 2$) was also calculated to see the magnitude of the difference between groups, following the standard rules (0.01 – small; 0.06 – medium, 0.14 – large) (Cohen 1988). To identify any mean differences between groups, multiple comparisons were carried out using Turkey post-hoc tests upon significant results from a two-way ANOVA.

Findings

This study explored the accuracy, consistency and existence of assessment bias in peer marking during *formative* and *summative regimes*. Accordingly, the two main effects tested are the effects of ACL and SPL on the performance score. In addition, the interaction effect to assess the prevalence of assessment bias, which is the combined effect of both ACL and SPL, is tested on the performance score. Based on the significance of the results of each of the effects, further analysis was conducted to gain more insight into differences between groups. The following section reports the results in four parts: testing of assumptions and descriptives, accuracy results; reliability results; and prevalence of assessment bias.

Testing of Assumptions and Descriptives

For the *formative regime* data, the test for normality results indicated acceptable values of skewness and kurtosis satisfying the normality assumption. The test for homogeneity of variance was significant for the formative scenario, with $p>0.01$ Levene $F(5, 92) = 2.5$, $p = 0.03$, indicating that homogeneity of variance assumption underlying the application of the two-way ANOVA was met. An alpha level of .05 was used for the analyses. For the *summative regime* data, the normality test results indicated moderate non-normality with skewness and kurtosis results slightly exceeding the threshold absolute values of 3. Since the ANOVA tolerates violations to normality assumption well, transformation of data to conform normality was not considered in this study. However, the homogeneity of variance assumption was violated with insignificant p-value (Levene $F(5, 92) = 18.5$, $p = 0.0$). Therefore, a Bonferroni adjustment was made to allow for more stringent tests of the differences in the average performance scores between groups. The Bonferroni adjustment is derived by by dividing the original critical P value by the number of comparisons made (i.e. 0.01/3=0.003) to get the new critical P value. Therefore using Bonferroni adjustment the minimum threshold level of significance was set at 0.003 to proceed with analysis.

Descriptive statistics consisting of means and standard deviations were used to describe the performance score measure for each group based on the ACL and SPL (Figure 6. For the overinflated group, mean scores are significantly higher in the *summative regime* compared to the *formative regime*, specifically for under and over contribution levels. However, the same is not true for the equal contribution level group. In contrast, for the accurate/modest groups, mean scores were significantly higher for the *summative regime* in comparison to the *formative regime* for all three levels of contribution.

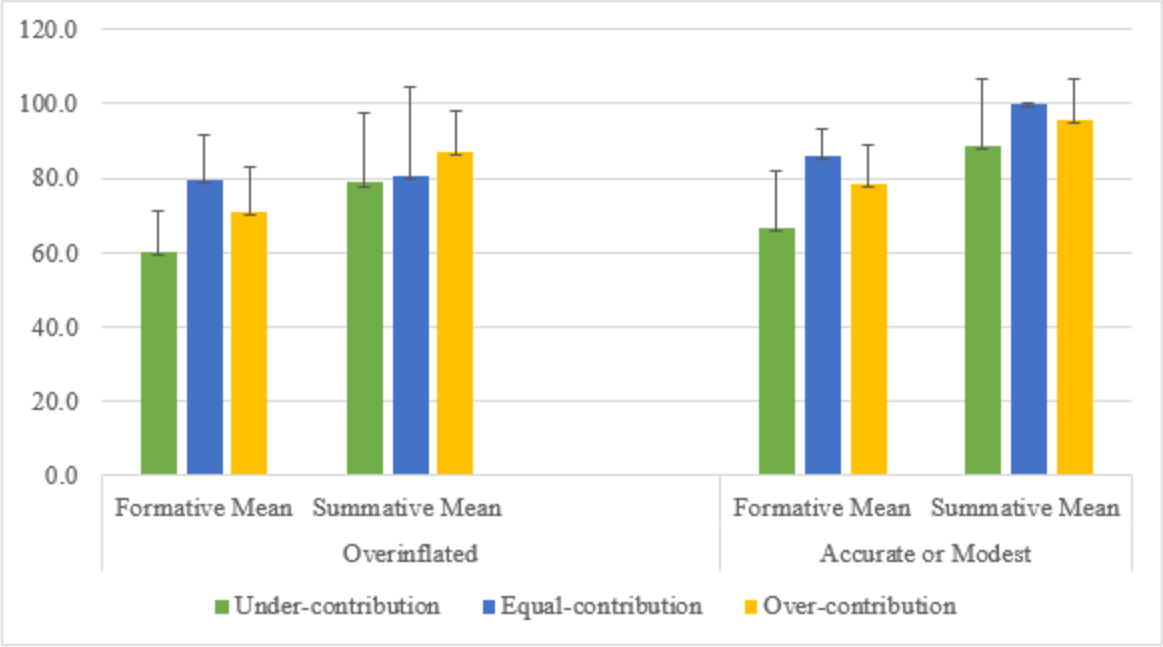

Fig.6 Comparison of Mean performance score by ACL and SPL

Accuracy

The accuracy of peer assessment was evaluated by comparison of performance scores based on ACL and SPL during both assessment regimes. Accordingly, a two-way ANOVA was conducted to evaluate the effects of the contribution levels and perception levels on performance scores under the formative and summative assessment regimes. This tested the mean performance score of students' group based on ACL (under, equal and over contribution) and SPL (overinflated and accurate/modest). If peer assessment was accurate, we expect the results to show significant results for main effects (for each of the variables) indicating that performance scores are discriminated based on the ACL and SPL classification for both *formative regime* and *summative regime*. The results of the two-way ANOVA in the *formative regime*, indicated a significant main effect for both ACL and SPL (see Table 2). The results suggest students with different ACLs showed significantly varying performance scores ($F(2, 92)=21.9$, $p=0.0$, $\eta 2 = 0.32$). Also, students with different SPLs, showed a significant difference in performance scores ($F(1, 92)=8.1$, $p=0.0$, $\eta 2 = 0.08$). The results indicate large effect size of 32.2% for ACL and 8.1% for SPL. This suggests that students' performance scores are differentiated based on the ACL and SPL.

Table 2. Performance score results by effects

Variables	Formative performance score					Summative performance score				
Effect	SS	df	MS	F	Effect Size	SS	df	MS	F	Effect Size
ACL	5432.2	2	2716.1	21.9**	0.322	453.9	2	227.0	1.8	0.0
SPL	1007.3	1	1007.3	8.1**	0.081	1790.3	1	1790.3	14.1*	0.1
ACL and SPL	2.2	2	1.1	0.0	0.000	304.3	2	152.2	1.2	0.0
R^2	0.39					0.35				
**P-value<0.01; *P-value<0.003										

With significant main effects for both ACL and SPL during the *formative regime*, further analysis was conducted to assess the accuracy of peer assessment using Tukey post hoc procedures to test the difference in mean scores. With respect to the comparison of results based on ACL, the results indicated that under-contributing student groups scored significantly less (MD=20.04, SD=2.8) in comparison to equal and over contributors (see Table 3). However, there was no significant difference in performance scores between equal and over contributors. With respect to comparison results for SPL, the results suggest that the overinflated level group scored significantly lower than accurate/modest group students (MD=-6.8, SD=1.6). These results suggest reasonable accuracy of peer marking during the *formative regime*.

Table 3. Comparison of formative peer assessment scores by level of contribution

Comparisons	Mean Difference Score	Std. Error	Lower Bound	Upper Bound
Under Vs Equal	-20.05*	2.88	-26.91	-13.19
Under Vs Over	-13.63*	2.70	-20.05	-7.21
Over Vs Equal	-6.42	2.75	-12.96	0.13

The results of the two-way ANOVA for the *summative regime* indicated a lack of the main effect for ACL and a significant main effect for SPL (see Table 2). This suggests there was no significant difference in performance score based on ACL ($F(2, 92)=1.8$, $p=0.2$, $\eta 2 = 0.0$). This indicates lack of accuracy in peer marking based on individual contribution. However, the significant difference in performance scores based on SPL ($F(1, 92)=14.1$, $p=0.001$, $\eta 2 = 0.1$), indicates accuracy of peer marking based on perceptions. More specifically, post hoc analysis based on SPL suggests that the overinflated level group scored significantly lower than accurate/modest group students (MD=-12.62, SD=3.4).

Consistency

The consistency of peer marking is evaluated by comparing the hypothesis results between the two assessment regimes (Table 4). If consistently assessed, we expect similar significant (or non-significant) results for both *formative and summative assessment regimes*. With respect to the main effect for ACL, the results indicate a significant difference in performance scores for the *formative regime* and, on the contrary, a non-significant

difference for the *summative regime*. This suggests lack of consistency in peer marking based on contribution levels. With respect to the main effect for SPL, the results are significant for both assessment regimes, indicating consistency in peer marking based on perceptions. Similarly, consistent marking is observed based on the lack of interaction effect during both *formative* and *summative regimes*.

Table 4. Hypotheses results for formative and summative regimes

Hypotheses	**Formative regime – support for hypotheses? (H1a, H2a, H3a)**	**Summative regime-support for hypotheses? (H1b, H2b, H3b)**
H1: Different ACLs will lead to varying performance scores	Yes	No
H2: Different SPLs will lead to varying performance scores	Yes.	Yes.
H3: Interdependency between ACL and SPL will influence the performance score	No.	No.

Assessment Bias

The existence of assessment bias was evaluated by exploring the interaction effect between ACL and SPL and its effect on performance score during both assessment regimes. The underlying premise for conducting the interaction analysis was that actual contribution and perceived contribution are intertwined in such a way that one may influence the other. The ANOVA results suggest lack of interaction effects during both *formative and summative regimes* between ACL and SPL. This indicates that there was no dependency/interaction between ACL and SPL. In other words, surprisingly, the overall results indicate the absence of assessment bias such as sucker effect or cognitive bias (halo effect) or friendly marking in peer assessment during both assessment regimes. However, breaking down the analysis of interaction effect between groups provides (Figure 7) more insight into prevalence of assessment bias during summative assessment, particularly for low performing overinflated group. This micro analysis is crucial in identifying underlying problem areas so as to adopt appropriate strategies that focus on specific problematic areas and groups.

Discussion

The objective of this study was to evaluate students' ability and behaviour in marking their peers in a collaborative group assessment when the overall mark is either counted or not counted towards the final grade. Our findings are encouraging and at the same time pose some challenges in using peer assessment for summative purposes.

Overall, the results suggest that students have the ability to mark their peers with reasonable accuracy, consistency and without bias when the mark does not count towards the final grade. However, in the *summative regime*, students' behaviour in marking their peers shows a dramatic shift with grade inflation and failure to differentiate high contributing students and their counterparts (see Figure 7). At the same time, perception (measured by SPL) has remained the same during both regimes with minor variation between the two groups.

To evaluate more specific differences, comparison between groups based on both variables and both regimes are discussed below.

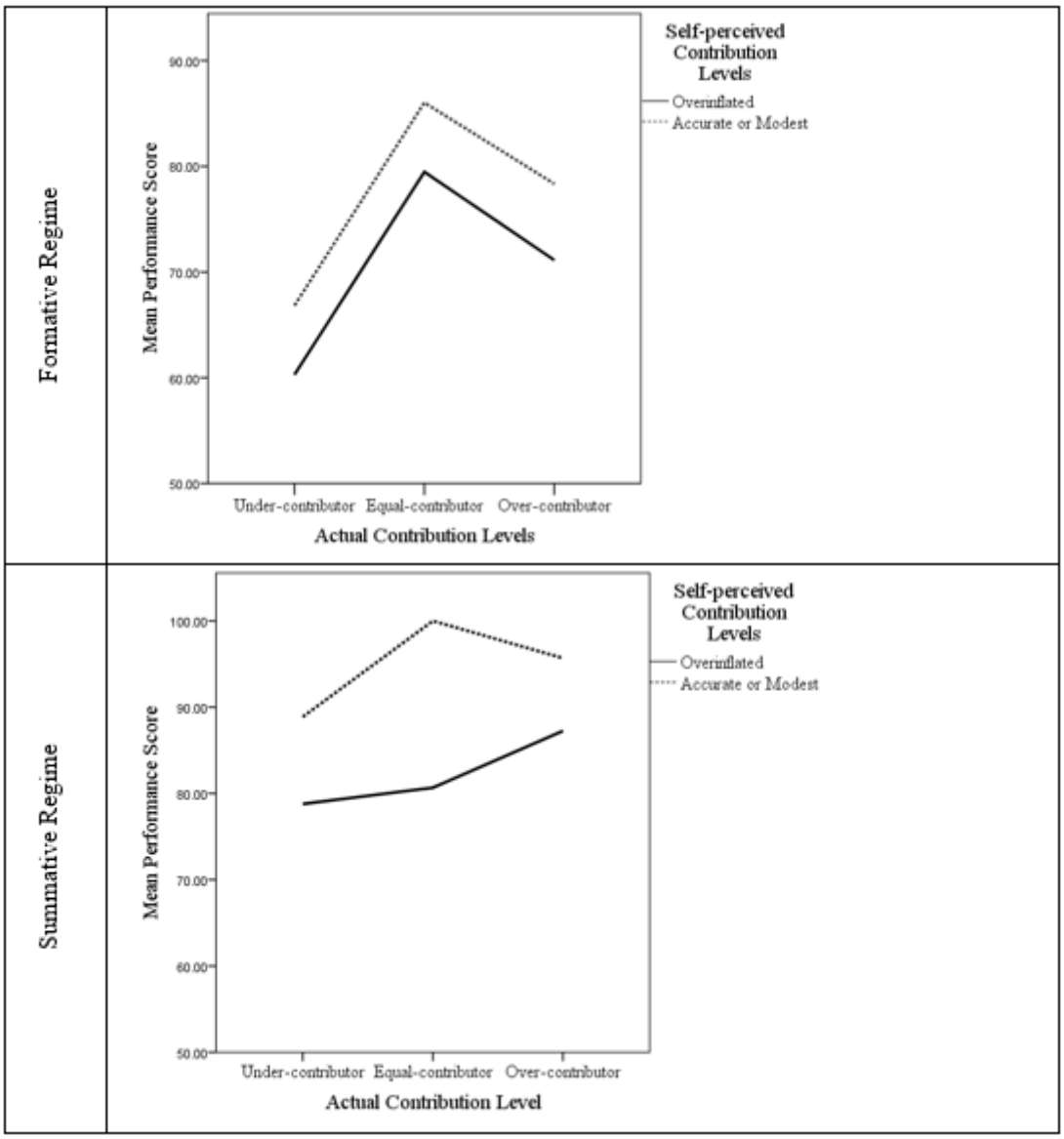

Fig.7 Comparison between performance mean scores based on ACL and SPL

For both groups of students, the *formative regime* results indicate accuracy in peer marking with a significant difference in mean scores, with under-contributors getting significantly lower scores than equal and over-contributors. These findings are in line with studies that demonstrate that students are capable of providing reliable and valid ratings (Cho, Schunn, and Wilson 2006; Falchikov and Goldfinch 2000). The results also suggest that low performing students tend to have overinflated views and high performing students having modest/accurate views about their contribution. These are consistent with results found in Boud and Falchikov (1989) with low performing students' inability and high performing students' superior ability to effectively and accurately self-assess themselves. The results in the *summative regime* indicate a lack of accuracy in peer marking with no significant difference in mean scores between under, equal and over-contributors. This

suggests that students, when confronted with inappropriate incentives, fail to grade accurately, which manifests itself in grade inflation when the assessment is counted towards the final grade. These findings are in line with friendly marking behaviour patterns during summative assessment scenarios (Steverding, Tyler, and Sexton 2016). This could be due to its high consequential impact on the final individual product mark in this study. Other underlying inhibiting factors that could contribute to such marking behaviour are: doubts about their capability to mark fairly, anxiety of retaliation and confrontation, and fear of the instructor breaking confidentiality.

Another key finding is no significance in mean score between equal-contributor and over-contributor during both regimes. This suggests students are somewhat comfortable in penalising high performing peers, if they 'hog' or do not equally share their workload (or if they are over-enthusiastic, demonstrate controlling behaviours or do more than their share of their work). This is quite a rational behaviour and suggests that students recognise teamwork is about everyone contributing equitably. These outcomes are similar to findings which point to harsh marking for high performing students (Steverding, Tyler, and Sexton 2016) and underrating of more capable students (Brehm and Festinger 1957). Effective assessment design that encourages shared responsibilities and discourages competive and lone wolf behaviour could combat these problems.

Mixed results were found for consistency results by comparison between groups, similar to ANOVA results. Consistency in marking behaviour is noted with similar patterns for the accurate/moderate group for both regimes. However, lack of consistency in marking behaviour was found with a dissimilar pattern in the sign and direction of the line for the overinflated groups between the two assessment regimes. Scores were marginally higher for over contributors than for equal contributors (overinflated group). This again suggests that students are willing to overlook low-performing students inflating their contribution, on the proviso that they actually contribute. Students are also not willing to penalise their peers for poor performance, even though they are aware of it (revealed by the significant difference in scores based on the self-perceived contribution level). Along these lines, (Steverding, Tyler, and Sexton 2016) noticed that students were more generous in peer marking, particularly for low performing students. Lack of willingness to consistently mark peers, even though it could be perceived as empathic conduct, is counterproductive in supporting long term learning.

With respect to assessment bias, even though the overall ANOVA results indicated no significant assessment bias with lack of interaction between both variables (ACL and SPL) in both assessment regimes with insignificant interaction effects, comparison based on different groups suggests some level of assessment bias during the *summative regime*. The graph for the *formative regime* clearly shows no prevalence of assessment bias with parallel lines of mean performance scores for both groups (overinflated group and accurate/modest groups). However, slight assessment bias is noted during the *summative regime* with a wider gap and non-parallel lines between overinflated and accurate/modest groups. This suggests that peer marking behaviour is different for the two groups (accurate/modest and overinflated group) during the summative assessment regime. This was more pronounced for the equal-contributor group with wider variation marking between the two groups. This again reinforces the conclusion that low performing students with relative high contributions are marked more leniently in comparison to their counterparts.

Reliability and validity concerns regarding self and peer assessment are hardly new in a higher education peer assessment setting (Falchikov and Goldfinch 2000; Yao-Ting et al. 2010). Similarly to our results, which challenge the use of peer marking for summative purposes, (Steverding, Tyler, and Sexton 2016)

indicated inconsistency in peer marking with more divergence in marking between peers and experts. However, it is well documented that reliability and validity of marking is a big problem even among experienced assessors (Bloxham et al. 2016). Vickerman (2009) highlights that when consistent marking is difficult for teachers, it should not be surprising that students encounter similar difficulties in assessing peers. Andrade and Du (2007) point to the tension between students' and teachers' standards of quality work. Rust (2002) found that even where there were written and verbal briefings, students still differed in their understanding of some criteria in comparison with both their peers and their tutors. Therefore, is it unreasonable to expect students to be consistent or accurate in their evaluative judgement about their peers - who are less experienced and do not have the necessary skills or exposure or experience (Sadler 2010; Boud, Lawson, and Thompson 2013; Andrade and Du 2007).

In summary, the use of summative assessment in collaborative group assessment inhibits good judgement. On the one hand, empathetic marking behaviour due to the fear of compromising the relationship appears to lead to higher peer marking. On the other, the competitive assessment cultures heightened by summative assessments lead to lower peer marking (particularly for high performing students). This would be counterproductive in implementing honest peer marking and development of teamwork skills in collaborative learning environments. Besides, the distortion effects on behaviour of having peer marks used summatively might not be ameliorated in a high competitive context. Paradoxically, we are at the crossroads where summative assessments are required for evidencing requirements which conflict with the authentic development of collaborative teamwork skills.

These findings indicate a need for appropriate policy measures to induce a positive behavioural change in peer marking. Many recommendations and strategies for developing self-assessment skills (Boud 1989) and overcoming problems associated with collaborative group work (Davies 2009) could be adapted for both self and peer assessment scenarios. Such measures include training, moderation, competency demonstration, incentives, penalties and others. For instance, training measures could include capacity building of students in developing peer marking skills, training and practices to reduce differences in marking expectations, provision of detailed criterion-referenced rubrics explicitly stating the expectations and standards accompanied by discussion and examples of work in which the standards are embodied (Bloxham et al. 2016) and provision of marking guides for different levels of performance. There is also a need for more focussed and disparate practice, training and competency development sessions for low (step up their expectation) and high (lower their expectation) performing students to adjust their expectations. Other measures to improve honest marking comprise an implementation of a confidentiality agreement for peer marking, communicating the consequences for breach of agreements and teachers editing students' qualitative feedback if they suspect a student's identity will be revealed.

Our study contained both strengths and limitations. The strength was that it was a study undertaken in the context of normal course units with their usual assessment practices. One of the key limitations is the generalisability of results due to the moderate sample size situated within a specific teaching domain. Even though the sample size was deemed acceptable, a larger sample size would have facilitated the use of more powerful analyses. We did not differentiate between of the type of students (undergraduate vs postgraduate) in the study design, which may have impacted the sensitivity of our results. Additionally, we did not include a control group to verify the effects of formative assessment. Future research directions could also include

qualitative research around testing of assumptions of low performing students' self-assessment capabilities and behaviours.

This study has implications for both practice and research. The results help in making more effective peer assessment design decisions and enhancing teaching and learning practices. Learning activities could be modified to encourage collaboration rather than competition (incentives for encouraging peers to contribute). Multiple formative opportunities could be provided to enhance confidence in self and peer assessment. Other learning activities such as co-creation or exploration of criterion-referenced rubrics along with discussion and examples can also enhance shared understanding of standards and improved consistency in marking. Educating students on how their enabling grade inflation behaviour could hurt their peers in the long run may also reduce inaccurate marking. In terms of mark calculation, final mark moderation methods could be instituted to discourage students' discomfort around impacting on final mark. Allocating sufficient weight to the peer mark may also ensure students take this activity seriously. Future research might therefore investigate which combination of these types of modifications result in more accurate peer assessments and students' formation of good judgements.

Conclusion

This study examined students' judgement ability and behaviour in collaborative group assessment (teamwork) context by testing the relationship between actual and perceived contribution levels and performance outcomes. The results confirm that use of summative assessment inhibits good judgement, even though students are capable of judging peers' performance accurately, consistently and without bias when the mark is not counted towards the final grade. During a *formative regime*, students were more honest in assessing their peers, but tended to be overly generous when their mark is counted towards their final grade. This is more pronounced for underperformers in a summative situation. Some level of assessment bias was noted particularly for the *summative regime* and for low performing students.

This study contributes to the peer assessment literature in a number of ways. Firstly, it offers a novel method for exploring peer assessment accuracy, consistency and assessment bias in process context - especially when collaborative group assessment is a social process (and students are the most appropriate candidates to judge peers interpersonal skills and behaviours). It employs actual and self-perceived contribution to investigate peer assessment concerns. Furthermore, it compares peer marking behaviour patterns during formative and summative assessment regimes to address concerns regarding peer assessment. Finally, this study examines the interplay between actual and self-perceived contribution levels in assessing the prevalence of peer assessment bias. Work within this realm is important to ensure students are appropriately prepared for the future world of work, where teamwork is an important skill not only to possess, but to accurately determine others' ability to work in teams also.

Acknowledgements
This work was supported by funding from the Centre for Research in Assessment and Digital Learning (CRADLE), Deakin University, [grant number 2016/01]. We would like to thank Jamie Mustard for his input and comments and Jade McKay for her research assistance in preparing this paper.

References

Andrade, H., and Y. Du. 2007. "Student responses to criteria-referenced self-assessment." *Assessment & Evaluation in Higher Education* 32 (2):159-81.

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 & Evaluation in Higher Education* 41 (3):466-81.

Bloxham, S., J.d. Hudson, B. Outer, and M. Price. 2015. " External peer review of assessment: an effective approach to verifying standards? ." *Higher Education Research & Development* 34 (6):1069-82.

Boud, D. 1989. "The role of self-assessment in student grading." *Assessment in Higher Education* 14 (1):20-30.

Boud, D., R. Cohen, and J. Sampson. 1999. "Peer learning and assessment." *Assessment & Evaluation in Higher Education* 24 (4):413-26.

Boud, D., and N. Falchikov. 1989. "Quantitative studies of student self-assessment in higher education: A critical analysis of findings." *Higher Education* 18 (5):529-49.

Boud, D., R. Lawson, and D.G. Thompson. 2013. "Does student engagement in self-assessment calibrate their judgement over time?" *Assessment & Evaluation in Higher Education* 38 (8):941-56.

Brehm, J., and L. Festinger. 1957. "Pressures toward uniformity of performance in groups." *Human Relations* 10 (1):85-91.

Brutus, S., M.B. Donia, and S. Ronen. 2013. "Can business students learn to evaluate better? Evidence from repeated exposure to a peer-evaluation system." *Academy of Management Learning & Education* 12 (1):18-31.

Cheng, W., and M. Warren. 2000. "Making a Difference: using peers to assess individual students' contributions to a group project." *Teaching in Higher education* 5 (2):244-55.

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

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Davies, W.M. 2009. "Groupwork as a form of assessment: Common problems and recommended solutions." *Higher Education* 58 (4):563-84.

Deakin. 2014. "Live the Future: Agenda 2020." In. Geelong: Deakin University.

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.

Fellenz, M.R. 2006. "Toward fairness in assessing student groupwork: A protocol for peer evaluation of individual contributions." *Journal of Management Education* 30 (4):570-91.

GCA. 2015. "Graduate outlook 2014: Employers' Perspectives on Graduate, Recruitment in Australia." In. Melbourne, Australia: Graduate Careers Australia Ltd.

Gravetter, F.J., and L.B. Wallnau. 2005. "Essentials of statistics for the behavioral sciences." In. Belmont, CA: Thomson Learning Inc.

Holmes-Smith, P., E. Cunningham, and L. Coote. 2006. *Structural Equation Modelling: From the fundamentals to advanced topics*. Melbourne, Australia: Sream and Statsline.

Lejk, M., and M. Wyvill. 2001. "The effect of the inclusion of selfassessment with peer assessment of contributions to a group project: A quantitative study of secret and agreed assessments." *Assessment & Evaluation in Higher Education* 26 (6):551-61.

Loughry, M.L., M.W. Ohland, and D.J. Woehr. 2014. "Assessing teamwork skills for assurance of learning using CATME team tools." *Journal of Marketing Education* 36 (1):5-19.

Nicol, D., A. Thomson, and C. Breslin. 2014. "Rethinking feedback practices in higher education: a peer review perspective." *Assessment & Evaluation in Higher Education* 39 (1):102-22.

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

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):325-35.

Planas Lladó, A., L.F. Soley, R.M. Fraguell Sansbelló, G.A. Pujolras, J.P. Planella, N. Roura-Pascual, J.J. Suñol Martínez, and L.M. Moreno. 2014. "Student perceptions of peer assessment: an interdisciplinary study." *Assessment & Evaluation in Higher Education* 39 (5):592-610.

Rust, C. 2002. "The impact of assessment on student learning: how can the research literature practically help to inform the development of departmental assessment strategies and learner-centred assessment practices?" *Active learning in higher education* 3 (2):145-58.

Sadler, D.R. 2010. "Beyond feedback: Developing student capability in complex appraisal." *Assessment & Evaluation in Higher Education* 35 (5):535-50.

Sambell, K., L. McDowell, and C. Montgomery. 2013. *Assessment for learning in higher education*: Routledge.

Speyer, R., W. Pilz, J. Van Der Kruis, and J.W. Brunings. 2011. "Reliability and validity of student peer assessment in medical education: a systematic review." *Medical Teacher* 33 (11):e572-e85.

Sridharan, B., M.B. Muttakin, and D.G. Mihret. 2018. "Students' perceptions of peer assessment effectiveness: an explorative study." *Accounting Education*:1-27.

Steverding, D., K.M. Tyler, and D.W. Sexton. 2016. "Evaluation of marking of peer marking in oral presentation." *Perspectives on medical education* 5 (2):103-7.

Sung-Seok, K. 2014. "Peer assessment in group projects accounting for assessor reliability by an iterative method." *Teaching in Higher education* 19 (3):301-14.

Tabachnick, B.G., L.S. Fidell, and S.J. Osterlind. 2001. *Using multivariate statistics*. Boston, MA: Pearson Education, Inc.

Tai, J.H.-M., B.J. Canny, T.P. Haines, and E.K. Molloy. 2016. "The role of peer-assisted learning in building evaluative judgement: opportunities in clinical medical education." *Advances in Health Sciences Education* 21 (3):659-76.

To, J., and D. Carless. 2016. "Making productive use of exemplars: Peer discussion and teacher guidance for positive transfer of strategies." *Journal of Further and Higher Education* 40 (6):746-64.

Topping, K.J. 2009. "Peer Assessment." *Theory Into Practice* 48 (1):20-7.

Vickerman, P. 2009. "Student perspectives on formative peer assessment: an attempt to deepen learning?" *Assessment & Evaluation in Higher Education* 34 (2):221-30.

Wen, M.L., and C.-C. Tsai. 2006. "University students' perceptions of and attitudes toward (online) peer assessment." *Higher Education* 51 (1):27-44.

Willey, K., and A. Gardner. 2009. *SPARK plus: self & peer assessment resource kit user manual revision 1.7*. Sydney, Australia: University of Technology Sydney and University of Sydney.

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):429-43.

Yao-Ting, S., C. Kuo-En, C. Tzyy-Hua, and Y. Wen-Cheng. 2010. "How many heads are better than one? The reliability and validity of teenagers' self-and peer assessments." *Journal of Adolescence* 33 (1):135-45.