INTERNATIONAL CONFERENCE ON ENGINEERING AND PRODUCT DESIGN EDUCATION
10-11 SEPTEMBER 2020, VIA DESIGN, VIA UNIVERSITY COLLEGE, HERNING, DENMARK

SELF AND PEER EVALUATIONS OF STUDENT PERFORMANCE IN SKILLS BASED DESIGN PROJECTS

Derek COVILL[1], **Tim KATZ**[1] **and Steven SMITH**[2]
[1]School of Computing, Engineering & Mathematics, University of Brighton, United Kingdom
[2]School of Architecture and Design, University of Brighton, United Kingdom

ABSTRACT
Self and peer assessment have been shown to help improve student engagement in assessment tasks, to improve behaviour and maintain interest and attention levels, and ultimately to improve student performance. In this study, we critically evaluated the process, reliability, validity, benefits and drawbacks of self and peer assessment in a first-year product design project with the aim to also establish a recommended resolution for rubric grade boundaries. Students (n = 51) carried out five separate week-long design projects. They then undertook self and peer assessments and were assessed by the module tutors, by additional staff (both technical and academic not linked to the project module) and students on other years of the course. All staff had reasonable agreement with tutor grades, although there were mixed results on whether these were a valid means to accurately assess the project outcome – depending on which statistical analysis was adopted. For all staff, > 90% of their assessments were within 10% of the benchmark (agreed) grades, while for student self-assessments 84.2% were within 10% and for student peer assessments 60.5% of the overall grades were within 10% of the tutor's grades. When we analysed the data for individual tutor marks for all assessment criteria, 37.3% of grades were within 5% of each other, 85.8% were within 10% of each other, and the range within which 95% of grades could be bound was 20%, suggesting that the resolution of our rubric grade boundaries should be increased to 20% between grades boundaries for individual criteria in such practical assessment tasks in design. More generally, understanding the reliability and validity of assessment processes can help inform the assessment design to ensure an appropriate resolution is used for criteria.

Keywords: Self, peer, tutor, assessment, reliability, validity, precision, accuracy

1 INTRODUCTION

It is widely understood that engaging students in their own assessment tasks is good practice and that there should be more emphasis on creating assessment tasks of higher-level thinking and complex skills [1]. Furthermore, it has been shown that self and peer assessments can have positive effects on students' motivation to learn and overall performance [1].

In this study, we aim to critically evaluate the process, reliability (consistency of results), validity (accuracy of measurement of intended outcome), benefits and drawbacks of self and peer assessment in a first-year design project. We also draw comparisons with data published in the literature in related or parallel fields. It is our aim to engage students in the assessment process right from the outset on our course, but also our wider intention in this study was to critically evaluate our assessment processes. We scrutinised the quality and clarity of our assessment criteria and considered what would be an appropriate resolution between grade boundaries for tasks of this nature in order to improve our curriculum design and the delivery of our design courses in the future.

2 LITERATURE REVIEW

Self and peer assessment have been used extensively in the past in many education settings and generally the understanding is that this is a useful and positive learning exercise for students. It has been shown to help improve student engagement in assessment tasks and the criteria, to improve behaviour and maintain interest and attention levels, and ultimately to improve student performance [2]. While there are many positives associated with using self and peer assessment, the process can potentially increase

the assessment time and effort for staff and there is clear evidence to suggest that students' self-interests can lead them to inflate or tend towards the median (for weak students) or deflate (for strong students) assessments of their own work [3]. However, this effect has been shown to be considerably reduced or even eliminated by training students how to assess their work [2], or where students are involved in the development of the assessment criteria [3]. Ultimately, peer assessment has been shown to be of adequate reliability in a wide variety of applications [3].

Self and peer assessment has been used in primary/middle school [1,2], secondary school [4] and higher/tertiary education [3,5-7] and across many subject areas from engineering [6], language and writing courses [4,8], management [9] and physiotherapy [10] and can be used to add value to existing assessment with tutors or as a substitute to staff assessment, although in the latter scenario quality checks typically remain in place [3]. Various approaches have been used to assess the correlation, reliability and validity of self and peer assessment methods including to assess the proportion of students who have assessed within 5% or 10% of the staff grades or mean differences from staff grades [11], to use statistics such as Student t-tests to analyse the effects of gender on self-assessment scores [8], analysis of variance (ANOVA) to assess the influence of age on self-assessment scores [8], and the assessment of the relationship between self-assessment scores and test scores by tutors using the correlation co-efficient, *r*, also known as Pearson product-moment correlation [3,8]. Evaluations of self and peer assessment methods has been used with student numbers ranging in orders of magnitude from the 10's (this study), to the 100's [9], to the 1000's [11]. In a study with various comparisons of reliability [10], a higher correlation was recorded between peer assessments of a piece of work ($r = 0.85$) when compared to those between self-assessment and peer assessment ($r = 0.64$). Furthermore, correlations between tutor and self-assessments were even lower ($r = 0.21$), and those between tutor and peer assessment were as low as $r = 0.34$ [10]. There appears little in the literature relating to the reliability and validity of assessment methods relating to practical design skills, however in the subject area of product design, assessments of subjective aspects of design projects (e.g. form or aesthetics) have been shown to have variations as high as 75% across four independent assessors [7] warranting further investigation in this area.

3 METHODS

First year product design students ($n = 51$) were required to undertake a series of week-long design projects culminating in five separate prototype models which were exhibited in our design studio. This compulsory induction project for all new students covers some key design principles and is a chance to level the playing field as students often come from a variety of artistic or technical backgrounds. Students were provided with a detailed assessment proforma which included an assessment rubric, highlighting the grade descriptors and associated marking grades for each criterion, with a space for comment. For each of the five projects there were two criteria relating to a) the design development – assessing the depth of understanding demonstrated through their development portfolio, and b) the design communication – assessing the how professionally the design intent had been communicated through the use of a physical model and also a marker render and/or CAD model. This meant that for each assessment there were 5 x 2 = 10 individual grades allocated, with an overall mark resulting from the average of these 10 grades since the criteria were equally weighted.

Students were made aware that their work was to be assessed by the module tutors who set the assessment task (tutors) and potentially a number of other assessors including: themselves (self), a number of peers from the same cohort (peers), as well as a number of peers from the second year on the course who carried out the same set of projects the previous year (other peers), the main technician who supported all of the design projects in the workshops (technician) and finally an independent member of academic staff with over 15 years of experience tutoring design projects who was not involved in the project at all, only attending the assessment day at the end (external tutor). Each student was asked to carry out two peer assessments on the students placed either side of them in the exhibition. The external peers ($n = 8$), technician ($n = 11$) and external tutor ($n = 11$) were required to assess only a randomised subset of the group. Prior to a joint assessment of all students, the module tutors both also independently assessed a larger randomised subset of just over half the group ($n = 26$) allowing for a direct comparison between the academic tutors who had close contact with the students throughout the projects. Their joint/agreed grades for all students ($n = 51$) was taken as the benchmark against which all other assessments were compared.

In terms of statistical analysis, we used two approaches to make judgements on whether self, peer and tutor assessments were valid or reliable. For all assessors we analysed the proportion of assessment grades which were within 5% or 10% of the agreed tutor grades (as used by [11). If the proportion of overall grades that were within 10% of the tutor grade was >= 95% of the total number of grades, then the method was considered to be valid (for tutor/self and tutor/peer comparisons) or reliable (for tutor/tutor, self/peer, and peer/peer comparisons).

The second method used was the Pearson product-moment correlation coefficient (critical value, r) using a two-tailed Student T-test with statistical significance at $p < 0.05$ to determine if there were statistically significant relationships between the grades generated by the various marker groups as per [1, 3, 8, 10, 11]. If there were significant relationships between the grades, then we judged that as indicating validity (for tutor/self and tutor/peer comparisons) or reliability (for tutor/tutor, self/peer, and peer/peer comparisons).

The project was approved locally by our school research ethics committee, and while it was mandatory for students to participate in the standard (tutor) assessment process since this was a formal requirement of the course, students were able to opt in to the self and peer assessments and if they did they were then given the further opportunity to opt in to the research project and were asked to provide written, informed consent for their anonymised data to be used as part of our analysis. This provided an opportunity for students to participate in the self and peer assessments even if they didn't want their data analysed as part of this study and all students were given the opportunity to withdraw their participation at any time without any consequences. All participating students were also asked to complete a short, voluntary, anonymous questionnaire which prompted for feedback on the overall experience, as well as the perceived value and pitfalls of self, peer and tutor assessment and feedback.

4 RESULTS

The proportion of grades with differences between assessors within 5% and 10% can be seen below in Table 1, along with a classification on whether the comparison was valid or reliable, which was considered the case if the proportion of overall grades was within 10% of the tutor grades more than 95% of the time. The tutors, technicians and external tutors all had > 90% of their assessments within 10% of the comparison grades, while the peer assessments had only 60.5% of the overall grades within 10% of the tutor's grades. Figure 1 shows an example of the comparison between the overall grades ($n = 26$) achieved by students when assessed by both module tutors independently (left), and the individual grades ($n = 260$) for all criteria also assessed by both module tutors independently (right) which has a considerably larger spread ($r = 0.65$ vs $r = 0.46$). Table 2 shows the Pearson product-moment correlation coefficient results in the comparison between the various assessor groups (for both overall student grade and individual criteria grades), highlighting those which have statistically significant relationships at both $p < 0.05$ and $p < 0.01$ levels, with only the tutors/external tutor comparison showing no significant relationship, or rather a statistically significant difference.

Table 1. Proportion of grades with differences within 5% and 10% and whether the comparison was valid or reliable (if proportion >= 95% within 10% of agreed grade)

Assessor comparison	***proportion of overall grades within 5%***	***proportion of overall grades within 10%***	***Classification***
Tutor A/Tutor B ($n = 26$)	63.4%	96.2%	Reliable – yes
Self/Peer ($n = 48$)	47.9%	77.1%	Reliable - no
Peer/Peer ($n = 10$)	60.0%	80.0%	Reliable - no
Tutors/Technician ($n = 11$)	81.8%	100.0%	Valid - yes
Tutors/External tutor ($n = 11$)	63.6%	90.9%	Valid - no
Tutors/External peers ($n = 8$)	71.4%	85.7%	Valid - no
Tutors/Self ($n = 38$)	39.5%	84.2%	Valid - no
Tutors/Peer ($n = 69$)	37.7%	60.5%	Valid - no

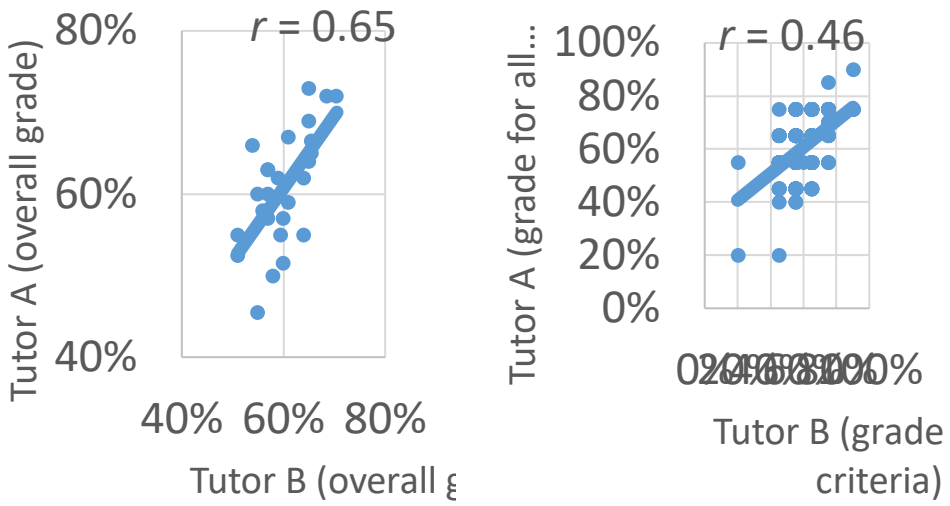

Figure 1. An example of the comparison of overall grades (n = 26) achieved by students when assessed by both module tutors independently (left), and the individual grades (n = 260) for all criteria also assessed by both module tutors independently – noting that many data points are coincident (right)

*Table 2. Pearson product-moment correlation coefficient or critical value, r_{value} - using a two-tailed tailed test with statistical significance at p < 0.05. The main values apply to the single/overall mean result for a student, with the bracketed values applying to all 10 separate grades for each separate criterion. These are compared against the benchmark r_{min} values. Note: * indicates significance also at p < 0.01*

Assessor comparison	$r_{min,\ p<0.05\ (0.01)}$	r_{value}	***df***	***Statistically significant relationship present?***
Tutor A/Tutor B (*n* = 26)	0.39 (0.20)	0.65 (0.46)	24(258)	Reliable - yes* (yes*)
Self/Peer (*n* = 48)	0.29 (0.20)	0.33 (0.33)	46 (478)	Reliable - yes (yes)
Peer/Peer (*n* = 10)	0.63 (0.20)	0.32 (0.26)	8 (98)	Reliable - no (yes*)
Tutors/Technician (*n* = 11)	0.60 (0.20)	0.70 (0.37)	9 (108)	Valid - yes (yes*)
Tutors/External tutor (*n* = 11)	0.60 (0.20)	0.45 (0.05)	9 (108)	Valid - no (no)
Tutors/External peers (*n* = 8)	0.75 (0.25)	0.83 (0.36)	5(68)	Valid - yes (yes*)
Tutors/Self (*n* = 38)	0.33 (0.20)	0.53 (0.50)	36 (378)	Valid - yes*(yes*)
Tutors/Peer (*n* = 69)	0.25 (0.20)	0.29 (0.27)	66 (678)	Valid - yes (yes*)

5 DISCUSSION

The results from the student questionnaire found that the majority of students found the exercise very or extremely useful (50%) and worthwhile (58.4%), with the majority of students (91.6%) stating that they would like self/peer assessment to be part of other assessments, but also found self-assessment more challenging (58.3%) than peer assessment, despite the fact that there was a highly significant relationship between their own assessments of themselves and those of the tutors, and a lower (but still significant) relationship between peer-peer assessments as shown in Table 2 above. Interestingly the correlation between tutors-self assessments were considerably stronger than those of [10], while the peer-peer and self-peer assessments in our study were considerably lower than those of [10]. We believe that this is due to students in our study not wanting to be overly harsh on their peers yet felt more comfortable being realistic with assessments of their own work. Our students were new to higher education, studying in their first year, and indeed this was the first set of design activities undertaken, and we believe this has influenced their peer assessments accordingly. We have extracted the data from Table 2 above and drawn a direct comparison with the results of [10]. This comparison is shown below in Table 3.

Table 3. A comparison of Pearson product-moment correlation coefficients between this

*study and those presented in [10]. Critical value, r_{value} – was calculated using a two-tailed tailed test with statistical significance at $p < 0.05$. The main values apply to the single/overall mean result for a student, with the bracketed values applying to all 10 separate grades for each separate criterion. Note: * indicates significance also at $p < 0.01$*

Assessor comparison	***Equivalent r_{value} [10]***	***This study r_{value}***	***Statistically significant relationship present?***
Self/Peer (n = 48)	0.64	0.33 (0.33)	Reliable - yes (yes)
Peer/Peer (n = 10)	0.85	0.32 (0.26)	Reliable - no (yes*)
Tutors/Self (n = 38)	0.21	0.53 (0.50)	Valid - yes*(yes*)
Tutors/Peer (n = 69)	0.34	0.29 (0.27)	Valid - yes (yes*)

Overall, the validity and reliability results appeared mixed depending on which statistical assessment was used. A clear limitation here is simply the number of assessments that were conducted; a much larger scale or repeated study would be ideal. In the first method used, only the assessments of the technician were valid when compared with the tutor grades, and only the tutor-tutor comparisons were reliable. However, in the second method, only the external tutor was considered invalid and the peer-peer assessments were not sufficiently reliable. Indeed, our definition of validity, being a comparison to the agreed tutor marks, was a relatively crude one which is a clear limitation in this study. Even so, it was surprising to see such large variations between assessments of individual criteria between staff, that the assessments carried out by the external tutor was significantly different from those of the module tutors. Furthermore, when comparing the overall grades given to students, i.e. the mean values from all 10 criteria, such discrepancies became less pronounced. As a result, we asked the following questions of ourselves: what would be a suitable resolution between grade boundaries? And have we sufficiently trained/informed the students in the early stages of the project on the self and peer assessment process? In the current study, the resolution of the scale included marks allocated as either 20%, 45%, 55%, 65%, 75% and 90%, and upon reflection this seems excessive, with module tutors achieving within 10% of each other for over 95% of their overall grades, and this is the absolute best case scenario. When we analysed the data for these individual marks for all criteria, only 37.3% of grades were within 5% of each other, only 85.8% were within 10% of each other, and the range within which 95% of grades could be bound was 20% (expected to be higher for self-peer assessments). It follows an appropriate resolution for our rubric should be in the order of 20% (or possibly greater) between grades boundaries which would assign the grade boundaries into 0-20-40-60-80-100 bands equivalent to fail-refer-pass-merit-distinction and this has informed our assessment practices for later projects. Furthermore, in this study we have not analysed whether the lack of accuracy is consistent across the range of grade boundaries, with the accuracy potentially varying non-linearly along the scale. While not covered within this study, it could be beneficial to consider the effect that the definition/clarity of the criteria used had on the learning experience, validity and reliability of the measurements used. Furthermore, we believe that engaging students not only in a training exercise to train them how to carry out the self and peer assessment exercise (as suggested by [2]), but to engage them actively through discussion and negotiation in the design of the assessment criteria is a hugely beneficial exercise and one that could further improve student engagement with the exercise and improve reliability and validity of the self-peer assessment results. While other studies have suggested that there can be a relationship between student capabilities, suggesting that weak students tend to inflate and strong students deflate their own marks [2,3], we have not evaluated this as part of this study, although we suspect from anecdotal evidence that this trend might exist. As such, further work will be for us to carry out further training with students, with mock assessments on previous submissions and build this exercise into them being actively involved in the design of the criteria as suggested above.

6 CONCLUSIONS

In this study, we critically evaluated the process, reliability, validity, benefits and drawbacks of self and peer assessment in a series of first year design projects that form part of a practical induction module with a comparison to key findings from the literature. The tutors, technicians and external tutors all had reasonable agreement with tutor grades, although there were mixed results on whether these were valid means to accurately measure the project outcome – depending on which statistical analysis was adopted.

When we analysed the data for these individual tutor marks for all assessment criteria, the range helped us identify what would be an appropriate resolution for grade boundary scales for the tasks in order to improve our curriculum design and the delivery of our design courses in the future. Our next steps for this research are to work more closely with the students to engage them in co-designing their own assessment criteria and embed more detailed training including mock assessments of work from previous years. Understanding the reliability and validity of assessment processes can help inform the assessment design to ensure an appropriate resolution is used for criteria.

REFERENCES

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

[2] Ross J.A. The reliability, validity, and utility of self-assessment, *Practical assessment research and evaluation*, 11(10), 2006, 1-13.

[3] Topping K. Peer Assessment Between Students in Colleges and Universities, *Review of Educational Research*, 1998, 68(3), 249-276.

[4] Chong I. How students' ability levels influence the relevance and accuracy of their feedback to peers: a case study. *Assessment writing*, 2017, 31, 13-23.

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

[6] Humphries P. Lo V. Chan F. Duggan G. Developing Transferable Groupwork Skills for Engineering Students, *Int. Journal of engineering education*, 2001, 17(1), 59-66.

[7] Pritchard G. Albon R. Objective assessment in Product Design education- Addressing the issue of marker.
variance, *Evaluations and Assessment Conference*, 2001

[8] Mistar J. A study of the validity and reliability of self-assessment, TEFLIN, 22(1), 2011, 45-58.

[9] Howlett D. Daruwalla P. Carter L. Davies D. Fisher G. & Hughes R. The role of criteria in peer assessments in management education and development: implications, considerations and findings,
Proceedings of the 19th ANZAM Conference, 2005.

[10] Lennon S. Correlations between tutor, peer and self-assessments of second year physiotherapy students in movement studies. In S. Griffiths, K. Houston, & A. Lazenblatt (Eds.), *Enhancing student learning through peer tutoring in higher education*: Vol. 1, 1995, 66-71.

[11] Kulkarni C. Wei K. Le H. Chia D. Papadopoulos K. Cheng J. Koller D. Klemmer S. Peer and Self-Assessment in Massive Online Classes. *ACM Transactions on Computer-Human Interaction*, 2013, 20(6), 33:1-31.