Cite as: Strang, K. D. (2015). Effectiveness of peer assessment in a professionalism course using an online workshop. *Journal of Information Technology Education: Innovations in Practice, 14,* 1-16. Retrieved from http://www.jite.org/documents/Vol14/JITEv14IIPp001-016Strang0350.pdf

Effectiveness of Peer Assessment in a Professionalism Course Using an Online Workshop

Kenneth David Strang
State University of New York, College at Plattsburgh,
School of Business & Economics, Queensbury, NY, USA

kenneth.strang@gmail.com

Abstract

An online Moodle Workshop was evaluated for peer assessment effectiveness. A quasi-experiment was designed using a Seminar in Professionalism course taught in face-to-face mode to undergraduate students across two campuses. The first goal was to determine if Moodle Workshop awarded a fair peer grader grade. The second objective was to estimate if students were consistent and reliable in performing their peer assessments. Statistical techniques were used to answer the research hypotheses. Although Workshop Moodle did not have a built-in measure for peer assessment validity, t-tests and reliability estimates were calculated to demonstrate that the grades were consistent with what faculty expected. Implications were asserted to improve teaching and recommendations were provided to enhance Moodle.

Keywords: Moodle Workshop, learning management system, peer assessment, reliability.

Introduction

Peer assessment can be a useful pedagogy, but it is not necessarily reliable, especially when completed by undergraduate students (Gielen, Dochy, & Onghena, 2011). If peer assessment was effective faculty could reduce their time spent on grading qualitative course assignments from large undergraduate classes, in addition to using this as pedagogy to improve student learning. Pairwise assessment consistency can be estimated by Kappa interrater reliability (Cohen, 1968). However, there were no best-practices for evaluating the reliability of multiple assessments–beyond two students although the literature contained several relevant studies. Zhang and Blakey (2012) used factor analysis for this while Dollisso and Koundinya (2011) applied t-tests.

Moodle is a popular open source Learning Management System (LMS); Moodle with its siblings Moodle Rooms and Joule are flexible solutions. Moodle Workshop (also referred to as Workshop further in the text) is the peer assessment module, but there are limitations according to practitioner experience (Mudrak, 2011b; Strang, 2013a). Most importantly, Workshop does not have a peer assessment reliability measure. More so, there is very little empirical literature for applying and validating Workshop for peer assessment.

Editor: Krassie Petrova

The goal of this study was to determine if Workshop could assign fair grades using peer assessment and, in parallel, to measure if students were consistent in their peer assessment when compared to faculty. The study sample was a face-to-face senior business course which had a large enrollment across two campuses. The lectures were classroom-based and all assessments were done in Moodle.

Literature Review and Research Hypotheses

Assessment of outcomes should be linked to learning objectives (Gielen, Dochy, Onghena, Struyven, & Smeets, 2011; Sadler, 2009). More importantly, assessment of learning should be accurate and relevant to student post-graduation needs (Biggs, 2003; Black & Wiliam, 1998). Two important issues that arise when designing assessment instruments are (1) who should be doing the assessing, and (2) how can the results be measured to ensure they are valid (Gibson & Dunning, 2012; Seery, Canty, & Phelan, 2012; Thomas, Martin, & Pleasants, 2011). It would be helpful to provide several basic definitions of key terms before exploring the empirical literature concerning peer assessment in a LMS.

The words assessment and evaluation are frequently used interchangeably in the literature, but they differ in significant ways (Bedore & OSullivan, 2011; Gielen, Dochy, & Onghena, 2011). Assessment results in written, oral, observational, and/or quantitative performance marks (e.g., test scores) that provide information to determine how well a student has progressed toward the intended objectives (Gielen, Dochy, Onghena et al., 2011; Green & Johnson, 2010). Evaluation uses the assessment to make judgments about a student's ability and to inform decisions about continued pedagogy (Gielen, Dochy, Onghena et al., 2011; Green & Johnson, 2010). Therefore, peer assessment is concerned with the grading of assignments by faculty or students based on predefined criteria while faculty will usually evaluate assessment scores to inform pedagogy.

The words formative and summative are often mentioned in the context of assessment. Formative assessment refers to a pedagogical process applied by the professor or students during the course to measure student understanding of the material, as well as to monitor and guide future pedagogy (Gielen, Dochy, & Onghena, 2011; Green & Johnson, 2010). Summative assessment is the evaluation done at the end of the teaching process for a group of concepts, albeit not necessarily at the end of the course (Russell & Airasian, 2012).

Usually formative assessment is completed by the professor through questions posed during the course while summative evaluation is done at the end of a learning unit through tests or assignments with predetermined grading rubrics. Peer student assessments are often summative in nature (Gielen, Dochy, Onghena et al., 2011; Green & Johnson, 2010), but they could be formative depending on the application. "By definition, all student work that contributes to course grades [is] summative. … Grades may be pressed into doing double duty: formative and summative" (Sadler, 2009, p. 808). In his study Sadler (2009) implied that formative and summative peer assessment was useful as far as it provided an extrinsic motivation and intrinsic self-efficacy. Another benefit argued in the current study is that peer assessment reduces faculty workload.

The key theoretical problems with assessment, including faculty-graded assessment, are reliability, validity, bias, and automation using technology. Peer assessment reliability refers to the degree that scores on the assessment are consistent and stable across multiple raters, namely students, faculty or combinations of both (Green & Johnson, 2010). The three common sources of error in peer assessment that decrease reliability are occasion (differences in time and context), items (some raters may not fully understand all criteria or perceive them differently), and scoring issues associated with bias related the relationships between students and their raters (Gielen, Dochy, Onghena et al., 2011).

A clear design using objective rubrics can reduce bias and improve the validity while statistical estimates such as interrater agreement can be generated to measure reliability (Black & Wiliam, 1998; Finn & Garner, 2011; Gibson & Dunning, 2012; Russell & Airasian, 2012). An LMS can be used for peer assessment in order to streamline peer assessment implementation, “off load” faculty work, and to improve the effectiveness of the process as well as to bolster student learning (Bitter & Legacy, 2008).

Peer assessment validity is the extent to which the instrument provides an accurate, representative, and relevant measure of student performance for its intended purpose (Green & Johnson, 2010). Construct-related rigor can be obtained by ensuring the rubric wording is clear. Content-related validity refers to measuring the correct objectives. Criterion-related validity refers to using relevant and easy to understand scoring scales for the raters use, such as “good versus bad” wording, and nominal or ordinal scales, e.g., Likert 1 to 10 ratings (Strang, 2013b).

The validity and reliability of peer assessment can be affected by differences between rater, rubric creator, and student socio-cultural background (Li & Lei-na, 2012; Mok, 2011; Shih, 2011). In fact, researchers have argued that there would be disagreement between raters regardless of whether they were students or faculty (Falchikov & Goldfinch, 2000; Schroeder, Scott, Tolson, Huang, & Lee, 2007). Thus, the pivotal issue driving this study was providing the peer assessment reliability measure that Moodle Workshop does not currently have.

The idea behind randomized allocation or peer evaluators is derived from the concept of the normal distribution because individual differences should average out when the sample size is large (Russell & Airasian, 2012). Therefore, if the number of peer raters was large enough the assessment would be fair. Some researchers argue that evaluator differences should reflect the real world workplace, so this is another argument supporting peer assessments (Goda & Reynolds, 2010).

Falchikov and Goldfinch (2000) acknowledged that faculty may not use peer assessment because they are afraid students will not be able to evaluate assignments reliably or that student marks will not be consistent with what faculty would do. Other researchers concurred with this (Bedore & OSullivan, 2011). Nevertheless, this is an effective learning strategy and pedagogy because students learn how to improve based on the feedback from a formative assessment perspective, and faculty may use the assessment scores as part of the course grading in a summative manner (Black & Wiliam, 1998; Gielen, Dochy, & Onghena, 2011).

Additionally, on the assumption that student assessing is done fairly, peer assessment should significantly reduce the load of the evaluation work especially when enrolment is large and when the assignments consist of long written reports (Gibson & Dunning, 2012). To test peer assessment validity, the first hypothesis was formulated.

H1: Student peer assessment scores will be valid at the 0.80 interrater agreement level.

Falchikov and Goldfinch (2000) performed a landmark meta-analysis of 48 empirical student peer assessment studies, finding that student evaluations of their peers were effective, with Pearson Product Moment Correlation r ranging from 0.14 to 0.99 (mean r was 0.69). They weighted the r calculation by sample size and number of comparisons made. Thus, larger cohorts would have a greater influence on their result. The nature of the subject matter in these studies was generally qualitative assignments which they described as “academic product and process” (Falchikov & Goldfinch, 2000, p. 310).

When Falchikov and Goldfinch examined 48 studies they found the “mean correlation over all studies was 0.69, indicating evidence of agreement between peer and teacher marks on average” (2000, p. 314). This finding indicated that a large portion of these were valid based on the benchmark of 0.70 for reliability (Hair, Black, Babin, Anderson, & Tatham, 2006). The corresponding interrater agreement benchmark set by Cohen (1968) was 0.80 which evaluates to a square root of 0.64 (Strang, 2009) for comparison to a Pearson Product Moment or Spearman correlation R (slightly below the 0.70 cited above). Therefore, R=0.64 (preferably 0.70) is the lowest correlation coefficient between peer assessments which should be accepted to consider that peer assessments were reliable.

Speyer, Pilz, Van Der Kruis and Brunings (2011) searched 2899 studies in the educational psychology literature for the period ending May 2010 to report the use of peer assessment as pedagogy. They concluded that peer assessment was widely used and it was an effective educational intervention to improve learning. Their advice for making peer assessment effective was to use an instrument linked to the learning objectives which has high reliability and validity which is in line with what other practitioners have recommended (Gielen, Dochy, Onghena et al., 2011; Sadler, 2009). In effect what they were recommending from empirical knowledge was to use a rubric to improve objectivity and to increase consistency between raters.

They found most peer assessment rubrics did not provide sufficient psychometric measures to ensure students were receiving a fair result. An important assertion they mentioned was “an instrument for educational purposes can only be justified by its sufficient reliability and validity as well as the discriminative and evaluative purposes of the assessment” (Speyer et al., 2011, p. 583). A limitation of their research was that they reviewed only 1% (28) of those studies in detail, which did not appear to conform to the systematic sampling methodology they planned. Unfortunately, no guidelines were given for benchmarks (e.g., mean acceptable consistency) or by way of methods and formulas to implement peer assessments. Furthermore they did not differentiate between formative versus summative assessment yet according to their discussion the latter was assumed.

Ng and Lai (2012) conducted a peer assessment experiment in an IT-related teacher education program at the Hong Kong Institute of Education in Peoples Republic of China using a Google-sites wiki (N=16). A difference in their experiment as compared to this study was that the teachers were mature and experienced with assessment. In their study, the students had to develop a rubric and apply it. Then students had to provide constructive comments to their peers in order to help improve the content of the learning materials. Students had to rate their own assignment as well as to evaluate their peer assignments.

They detected inconsistency in the student peer assessment process. Interestingly, they found some groups always gave higher scores to peer assignments, but graded their own work lower. Some groups gave higher grades for their own assignments as compared to peers while the opposite condition was observed for the other teams. Regrettably, they concluded “there was no observable evidence that assessment rubrics can serve as viable guidelines for evaluating wiki projects through either self-assessment or peer assessment” (Ng & Lai, 2012, p. 79). Notwithstanding their negative findings, it may be worth emphasizing that the nature of their course content was Internet programming that they pointed out was difficult to assess.

Shafaei and Nejati (2012) examined 59 undergraduate business education students, finding that self-determination significantly enhanced student commitment in the program. A key part of developing self-determination was to include peer assessment and constructive feedback mechanisms in the curriculum. Their recommendation for educational administrators was to mandate learning and assessment.

In their meta-analysis Falchikov and Goldfinch (2000) calculated the correlation R of academic product and process assessments as (0.75 with combined N=39 studies). The cause-effect coefficient of determination r for the peer assessments in the business discipline was 0.71 (N=11). They calculated an overall weighted effect size from 24 experimental studies to be 0.24, which is a large effect (Cohen, 1992; Cohen, Cohen, West, & Aiken, 2003). This indicates that empirical studies have shown student assessments of their peers to be effective in terms of consistency with faculty evaluations of the same assignment.

Surprisingly, they also found that correlations between student and faculty peer assessment of assignments did not increase as the number of students increased. The optimal number of raters for peer assessment based on meta-analysis research was 3-5; with more raters, consistency drops (Falchikov & Goldfinch, 2000). Interestingly, they found that the quality of student peer assessment did not significantly differ across disciplines or based on tenure of the student (time in the program, such as year 1 versus year 4).

Li (2011) evaluated peer assessment in a project management course (similar to this study) at a university in Georgia (USA). She analyzed student perceptions and outcomes of peer assessment effectiveness as pedagogy. She found that students in early learning development stages showed more learning gains than high achieving students. However, all students held positive attitudes towards their peer assessment experience. This indicates the peer evaluation process was effective as a formative assessment. Li, Liu, and Zhou (2012) conducted a follow up study on this data that confirmed the importance of peer feedback. Their approach was to use assessments during the course to help students self-regulate their learning and also as a mechanism for grading.

Nulty (2011) published a study whereby he recommended using peer assessment early in the students learning cycle. Additionally he cautioned against the disadvantages of using self-assessments due to bias. Goda and Reynolds (2010) conducted an evaluation of a US Military training program. They concluded that peer assessments were useful for program evaluation, and, therefore, this was a relevant technique to increase student learning.

Liu and Lee (2013) investigated peer observation and feedback on student learning during a psychology course in Taiwan. They determined that peer assessment was helpful to students, especially later in the course timeline. An important finding from their work was that students became better at peer assessment after practice. Therefore, an important implication would be requiring students first to complete a practice peer assessment.

Idowu and Esere (2010) interviewed 500 randomly-selected teachers in Nigeria to investigate which assessment techniques were most effective. They advocated a “wholistic assessment measure” which included “test and non-test techniques” such as peer assessments (Idowu & Esere, 2010, p. 342). This is relevant for this study in as far as the peer assessment rubric ought to reflect both qualitative and quantitative competencies, such as clear communications along with accurate budget calculations. Additionally, oral presentations and written reports should be assessed while diagrams and tables should be required in the assignments in order to accommodate the different learning styles of the student and the assignment evaluators.

Some faculty use peer assessments informally rather than as grading mechanisms. Heyman and Sailors (2011) found that traditional peer assessments helped students learn the material better. They also proposed an interesting approach to better understand the perceptions and learning styles between raters and peers by having students nominate their raters. However, this would be time consuming for large classes involving multiple assessments. An important concept arising from their study was to reinforce the idea of students practicing peer assessments. The findings from these studies suggest peer assessments are valuable to use on a formative and summative basis. Nevertheless, randomization does not guarantee immature students will conduct a fair peer assessment. Therefore, a faculty evaluation or previous course average ought to be used to ensure the student ratings are consistent with a subject matter expert. Thus, the second hypothesis was.

H2: Mean student peer ratings will be consistent with a faculty benchmark score.

Most empirical studies investigated above concluded or confirmed that peer assessments were useful in the sense that they helped students to learn. One study did not find that. However, only two of those studies used accepted statistical techniques to confirm the validity of student peer assessments. Since these studies used a two-pair grouping structure, their approaches were not comparable to this study because here more than two students were assigned to assess peer work. Additionally, none of the above studies evaluated using Moodle Workshop for peer assessment. The third hypothesis was formulated to answer a summative research question.

H3: Moodle Workshop will be effective for administering peer assessment grades.

Methods and Sample

A theory-dependent positivist philosophy was applied in this study. This philosophy consisted of a deductive literature review to inform the research questions, a valid assessment instrument design, and the application of quantitative analysis methods (Gill, Johnson, & Clark, 2010; Strang, 2013b). Since this study was designed to collect performance data, quantitative techniques were selected to test the hypotheses (Creswell, 2009).

Descriptive statistics, correlation, interrater reliability, and validity tests were applied at the 95% confidence level. SPSS version 14.1 was applied for the statistical tests while Moodle version 2.4 and Workshop version 2.0 were installed at SUNY for this quasi-experiment. Parametric normality checks and then t-tests were used to compare student versus professor rating consistency.

In terms of sampling method, natural intact convenience groups (existing classes) were used at the SUNY Plattsburgh and Queensbury campuses, a public comprehensive university located north of the state capital Albany NY (USA). The enrollment at this university was 6350 matriculated students. From that, 1050 of those were enrolled in the School of Business and Economics. At the time of writing, 350 were in the undergraduate Bachelor of Science in Business Administration (BSBA) program.

The sample size was 114 students spread over two campuses and three sections in adjacent semesters. The syllabus was identical and the same professor taught all sections. The mean age of the sample was 23 (SD=2.1) while females represented 59% of the class. The demographic factor and GPA estimates of the sample were similar to the university's business school population (based on z-score tests). The z-score is a test which can show that the mean of a sample is similar to a population, so generalizations may be made on analysis of the sample (which is the purpose of inductive research). As evidence of sample similarity to the university's business school population, the mean grade of a prerequisite course (advanced econometrics) was not significantly different between these students as compared to their cohort population mean (Z=1.71, [N1=321, N2=114], P>0.05). The mean grade was also similar between gender when partitioned by females versus males (F=4.1, [DF=1,113], P>0.05). Thus, it was asserted that the ability of these students was similar in comparison to others at this stage in the degree program (based on performance in a prerequisite course), and the ability of students in this experimental sample were similar between genders.

The course was Seminar in Professionalism. The written assessments included four components: career plan, biography, cover letter, and resume. These were submitted periodically during the course at approximately equal intervals. Each item needed to be one page in length. Peer assessment was the fifth graded element. The course work lasted approximately thirteen weeks.

The grade for each assessment component was weighted evenly at 20%. All assignments were submitted into Moodle Workshop. The scores for the four written assessments were calculated as the un-weighted average of all peer generated scores. The grade for the peer assessment component was calculated by Moodle Workshop using the best assessment algorithm.

Students were randomly allocated five peer reviewers in Moodle Workshop. All peer reviews were based on a rubric (called aspect in Workshop), and each reviewer mark was weighted at 1.

The comparison of assessments of fair (2.5) was specified for all. The professor did not complete a review in Workshop; instead he manually assessed each assignment component using the rubric.

The Workshop module in Moodle is specifically designed to automate peer assessments. A grade is given for the assessment from peers, and a separate grade is given to each student rater. The grade for the assessment is simple - it is the average from all raters (with optional weighting if the instructor wishes to contribute a peer assessment). Assessments may be blinded in order that students do not know whom they are assessing – this was the setting applied here. Self-assessments are also possible, but this was not used in this study due to self-prophecy bias: Students will tend to overrate their own performance. Currently only positive integers (as Likert scales) are available in Moodle Workshop for ratings. This limits the applicable statistical techniques. There are two assessment formats: accumulative or rubric; each functions similarly.

At the time of writing, there was only one method implemented in Moodle Workshop version 2.0 for rater grading, which is called best assessment. The underlying methodology is not well explained, and a pilot study returned inconsistent results where two identical raters (having the same peer assessment scenarios) were given different scores. The basic idea is that the best assessment is identified, and the rater is given a coefficient based on the differences in their scores from the best one for each rubric aspect, ((best score - peer score) * weighting / max possible score))^2.

The Moodle 2.4 Workshop module version 2.0 documentation states:

> Grade for assessment tries to estimate the quality of assessments that the participant gave to the peers. This grade (also known as grading grade) is calculated by the artificial intelligence hidden within the Workshop module as it tries to do typical teachers job. There is not a single formula to describe the calculation. However, the process is deterministic. Workshop picks one of the assessments as the best one - that is closest to the mean of all assessments - and gives it 100% grade. Then it measures a distance of all other assessments from this best one and gives them the lower grade, the more different they are from the best (given that the best one represents a consensus of the majority of assessors). The parameter of the calculation is how strict we should be, that is how quickly the grades fall down if they differ from the best one. (Mudrak, 2011b, para 27)

The best assessment was determined for each rubric aspect based on finding the peer assessment grade from all raters that has a standard deviation very close to zero. “In some situations there might be two assessments with the same the variance (distance from the mean) but the different grade. In this situation, the module has to warn the teacher and ask her to assess the submission (so her assessment hopefully helps to decide) or give grades for assessment manually - there is a bug in the current version linked with this situation” (Mudrak, 2011a, para 2).

The grade for assessment (given to a student for assessing peers) is calculated using the comparison of assessments setting in Workshop. That field is then multiplied by the best assessment difference coefficient. The “comparison of assessments values are 5.00 = very strict, 3.00 = strict, 2.50 = fair, 1.67 = lax, 1.00 = very lax” (Mudrak, 2011a, para 3). For a simplistic example, if the best assessment difference coefficient were 10%, and if the fair setting were used for comparison of assessments, then the grade for assessment = 1- (10%*2.5) = 75%.

Results and Discussion

The student ratings for each of their five peer assessment scores were extracted from Workshop for analysis in SPSS. There were two ways to accomplish that. First, the MySQL tables may be accessed directly (not recommended). Secondly, the picture icons can be set to remain hidden and the screen from the Moodle Workshop may be pasted into an Excel spreadsheet for importing into SPSS. One column in the spreadsheet will be the mean grade. The peer assessment grades are in rows so these need to be transposed into additional fields beside the mean grade. The professor grade may then be added into the spreadsheet. These six variables along with a student identifier may be pasted from the spreadsheet directly into the SPSS data view (the variable types may have to be adjusted to metric in order to permit the parametric statistical techniques to be performed).

In this study, each assignment received five scores from grader1 through grader5. The average of the five ratings was assigned in Workshop as the Moodle grade. The professor added an independent assessment grade. The descriptive statistics are listed in Table 1. A Kolmogorov-Smirnov test was performed to test if all of the metric fields (grades) were normal; the p-value results ranged from 0 to 0.023 which confirmed all grades approximated a normal distribution.

A Skew and Kurtosis estimate close to or below absolute 1 shows the data are normal; that is a prerequisite for performing parametric statistical techniques (Hair et al., 2006). The assumptions for permitting further reliability analysis using parametric statistical techniques are that the data can be dichotomous, ordinal, interval, or ratio scales, and that the observations (rows) be independent of one another. All of these relevant prerequisites were met for the sample.

Table 1: Descriptive statistics of peer assessment scores

	Mean		Standard Deviation	Skew		Kurtosis	
	Average	Std. Error	SD	Skew	Std. Error	Kurt	Std. Error
Grader1	91.553	.8079	8.6258	-1.037	.226	.035	.449
Grader2	88.184	.8649	9.2342	-.316	.226	-1.002	.449
Grader3	91.474	.7379	7.8787	-.945	.226	-.060	.449
Grader4	91.140	.7947	8.4846	-.759	.226	-.615	.449
Grader5	89.623	.9822	10.4874	-1.251	.226	.950	.449
Moodle grade	90.395	.6947	7.4172	-.910	.226	.031	.449
Professor	91.147	.6185	6.6032	-1.109	.226	1.079	.449

Hypothesis Testing

In order to test the first hypothesis (student peer assessment scores will be valid at the 0.80 interrater agreement), it was necessary to determine the most appropriate method. Earlier studies cited in the literature review had used factor analysis and t-tests. T-tests are not appropriate for more than two graders. For this reason, many researchers use ANOVA. However, the problem with using ANOVA here is that it measures the variance across groups (all rows) and not the consistency between raters (within the row itself). ANOVA compares group means, but not different peer grades within a row of data. Factor analysis could be used, but again that approach is focused on identifying similarity between fields in a row over all rows, to identify a pattern of factors. It is clear that these statistical techniques are not appropriate for estimating the reliability of peer evaluations. Cohen (1968) developed Kappa as an interrater agreement formula. The problem, though, is that it applies to pairs of scores. It is not designed (in its current form) to work with more than two peer graders. SPSS can calculate the Kappa as a correlation agreement if the values are summarized into a contingency table, but this was not relevant for the current study since the dataset was in a raw format. A pilot study of this was conducted (Strang, 2013a), but the calculation was found to be very tedious to apply with more than a few rubric (aspect) items and student assignments. Reliability analysis was a logical approach since it was designed for this purpose.

Reliability analysis is a family of statistical techniques that allow a researcher to examine the properties of measurement scales and the items in a structure that composes those scales (Keppel & Wickens, 2004). These techniques estimate the relationships as coefficients between the individual items within a row or within columns of a dataset and generally aggregate the coefficients across all the rows or columns to form an overall indicator or index of reliability.

According to Strang (2009), the appropriate technique to explore the dimensionality of multiple variables beyond two in a dataset could include factor analysis or multidimensional scaling. These techniques identify homogeneous groups of factors while a related technique called hierarchical cluster analysis can visually group similar factors. As noted, Zhang and Blakey (2012) utilized factor analysis. However, there are simpler approaches discussed below.

Correlation is a basic technique which can estimate similarity as the strength of a relationship between two bivariate factors. Inter-item correlation coefficients can be used to compute inter-rater reliability estimates for a matrix of bivariate correlation coefficients.

Alpha reliability is a method to develop a model of internal consistency based on the average inter-item correlation between all the variables (Cronback & Snow, 1981), and more than two may be processed. Therefore, this is an excellent choice of technique for this purpose.

There are several alternative reliability approaches. The split-half approach requires dividing the dataset into two approximately equal parts and then comparing each using correlation – this is not appropriate since the goal is to compare between the rating (grades) in a row. Variations of this include the Guttmans, Parallel, and Strict Parallel models, but none of them would be appropriate.

Intraclass correlation is another approach that may be used to estimate the similarity and thereby reliability of values in a dataset. Intraclass coefficients include single and average measures. A single measure applies to a row such as the ratings of judges or students on the individual item scores, whereas average measure applies to the overall dataset, for example, the average rating for k students. This is an appropriate technique for this study.

Table 2 lists the inter-item (bivariate) correlation between each student grade (1-5) and the professor, with the Cronbach reliabilities on the diagonal. Inter-item correlations may be calculated as the mean of significant and non-significant calculations (Strang, 2009). In this study, the inter-item correlation coefficient for grades of all five students with one another and the professor was R=+0.649 with a range of bivariate correlation R from +0.465 to +0.825 (SD=0.008). Correlation R can be converted to approximate a Kappa interrater agreement by taking the square root, which would be 0.806 (acceptable according to Cohen, 1992). Additionally all of the Cronbach alpha reliabilities were significant and all over 0.70 as specified by (Hair et al., 2006).

Since all of the Cronbach Alpha reliabilities from Grader1 through Grader5 (bolded in Table 2) were above the benchmark of 0.70 (which is equivalent to an interrater agreement of 0.84), the first hypothesis (student peer assessment scores will be valid at the 0.80 interrater agreement), can be accepted. It appears that, overall, students were consistent with one another when rating (grading) the same assignments (per row in the dataset).

Table 2: Inter-Item Correlation Matrix (N=114)

	Professor	Grader1	Grader2	Grader3	Grader4	Grader5
Professor	**.757**					
Grader1	.724	**.909**				
Grader2	.699	.706	**.908**			
Grader3	.825	.653	.616	**.866**		
Grader4	.722	.589	.526	.699	**.880**	
Grader5	.640	.577	.465	.676	.621	**.906**

The average intraclass correlation coefficient of 0.908 is shown in Table 3 along with supporting descriptive statistics. This coefficient applies to the overall dataset, for example, the average rating for peer students on the assignment. As noted, interaction effects are not processed because it would be illogical to do so (different students could talk with one another about marking the same way but it would be difficult to detect this outside of surveying each person). The intraclass correlation coefficient of 0.908 is somewhat comparable to the Kappa interrater agreement. This provides additional support for the first hypothesis. This information is shown here because the technique is available in statistics software so it may be used in this way to confirm if students were consistent in their peer ratings when using Moodle Workshop.

This intraclass coefficient is the same calculation as the overall Cronbach alpha reliability. Therefore, the benchmark of 0.70 would apply. An additional estimate is available, namely the standardized Cronbach alpha reliability, which is calculated by considering the number of factors and variance. The Cronbach reliability for this dataset was 0.917, which supports the first hypothesis.

Table 3: Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.623[a]	.549	.696	10.906	113	565	.000
Average Measures	.908[c]	.880	.932	10.906	113	565	.000

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

The second hypothesis (mean student peer ratings will be consistent with a faculty benchmark score) was tested using a pair-wise t-test between the professor score and the mean grade. A pair-wise t-test was selected because there is a natural purposeful relationship between these two variables in the same row: they were intended to be estimates of the same student assignment. The t-test was conducted both with and without the assumption of equal variances between scores which produced the same results. First though, note that the correlation between peer grade and the professor score was +0.867 which was significant (P=0.000). This indicated that the two were strongly related as a whole. The pair-wise t-test looks at the data row by row.

The results of the pair-wise t-test (summarized in Table 4) were T(113)=-1.812 (P=0.073), which supported the second hypothesis that the student ratings were consistent with the professor. The low t value and p-value above 0.05 indicated there was no significant difference between the two columns of grades (student peers versus professor), which supports the second hypothesis.

Table 4: Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Grade - Professor	-.6298	3.7119	.3477	-1.3186	.05895	-1.812	113	.073

Finally, the third hypothesis (Moodle Workshop will be effective for administering peer assessment grades) required additional investigation and qualitative reflection to answer. First the researcher examined the rubric ratings to ensure Moodle Workshop properly calculated the average as well as the peer grader grades according to the best-assessment. A sample of 10% was selected (10% * 114 records = 12). The sampling method was random which systematic by selecting every 10th record. Checking the calculations was a labor intensive process since there are five results per single row to verify. The calculations were correct for all of the 12 rows that were examined which is 100% reliability. So mathematically, Moodle Workshop was valid and reliable.

The researcher reflected on the whole process of using Moodle Workshop for the peer assessment. Overall it was effective and efficient. Therefore, the third hypothesis was accepted in that Moodle Workshop was effective for administering the peer assessment process and grading it.

Study Limitations

The limitations of this study, beyond the small sample size of 114 students, are the context of the university, which may not generalize to other organizational cultures or practices. For example, other universities may not feel comfortable using Moodle Workshop, which forces a course into a digital infrastructure to some extent. A disadvantage related to this it the time it may take to initially learn how to use both Moodle Workshop and a statistical package like SPSS to conduct this methodology. Furthermore, this study used Moodle version 2 so it is unknown if the suggestions outlined here would be effective in newer versions of Joule or Moodle Rooms. This should be tested in future studies.

This study used a qualitative face-to-face course in the business discipline – Seminar in Professionalism – thus, the findings may not generalize to other disciplines or to pure online modalities.

Although the students were blinded from knowing who their peer was, those conducting the assessment would be aware of the identity because of the nature of the documentation (containing cover letters). Additionally while some faculty may be comfortable with having students perform peer assessment, they may reject doing this through an online LMS technology. In other situations faculty may not believe in the literature review cited in support having students conduct peer assessments. Finally, some universities may simply not have the ability to use or switch to Moodle LMS due to sunk costs and a change-adverse culture.

Conclusion

The purpose of this study was to investigate if Moodle Workshop was effective for peer assessment. The quasi-experiment was applied on a Seminar in Professionalism business discipline mandatory course taught in face-to-face mode with undergraduate students across two campuses (N=114). The first goal was to determine if Moodle Workshop would calculate a fair grader grade. The second objective was to measure if students were consistent with the professor in scoring the assignments. Nonparametric statistical techniques were used to test these hypotheses, including Kappa interrater agreement, Cronbach reliability, correlation and t-tests.

The results were that all three hypotheses were supported. Students were consistent in scoring the assignments as compare with one another (based on the Cronbach Alpha reliabilities from Grader1 through Grader5 being above the benchmark of 0.70 and intraclass correlation of 0.907). Students were consistent with the professor ratings on the same assignment, based on the results of a pair-wise t-test(113) = -1.812 (P=0.073), which indicated no significant differences in ratings. Finally, Moodle Workshop was considered effective and efficient based on a qualitative evaluation.

The observed benefits for faculty based on this study were:

- Sharing of peer assessment work;
- Creation and management of assignments in a digital e-portfolio type facility;
- Ability to force students to paste-in (to limit volume) or attach multiple files;
- Automatic switchover from submission to assessment mode on a certain date;
- Random allocation of five (or any number) of students to assess each assignment or instead to require five (or any number) of assessments per submission;
- Ability to allow students who did not submit to conduct a peer assessment.

However, there were two significant limitations or missing features in Moodle Workshop:

1. No Cronbach alpha reliabilities (but this study illustrated how that could be done);
2. Only one peer assessment methodology: best-assessment plug-in.

Students can learn from the peer assessment process, not only about how to assess, but they may also see alternative approaches for applying the theories taught in the course. Peer assessments were formative as well as summative in nature since they were distributed throughout the course schedule, and the scores contributed towards the final grades. Students appreciated the peer assessment pedagogy based on the fact that several made reflective comments in the course opinion survey. Students were very satisfied with this course, which had an overall mean rating of 4.5 out of 5 for the instructional items on the survey (SD=0.8, N=110 respondents).

The researcher noted the most significant benefit from this study was confirming the reliable application of the technology-enabled Moodle Workshop for peer assessments. Although the professor manually assessed every student assignment in this course (N=114), if the Cronbach alpha reliabilities had been available he could have just randomly sampled a few students to conserve a tremendous amount of time. This methodology would be extremely valuable for large cohorts in qualitative subject oriented courses where there are numerous items to assess.

There are additional implications for saving faculty time by sharing the assessment work. Peer assessment can remove the assessment burden from faculty. This is asserted because the current study has shown that a single reliability measure can be calculated to evaluate the quality of student peer assessment. The literature indicated peer assessment improves student learning. Another point is that conducting a peer assessment is not a particularly stimulating task for a professor, but students may relish in this new responsibility. Therefore, a professor's time could be better spent improving teaching materials, maintaining currency, and mentoring students.

It should be noted though that having students conduct peer assessments is contingent on the maturity level of the student cohort and dependent on the type of assessment instrument. For example, Workshop seems ideal for qualitative data such as written essays, but it may not work well for verbal/physical observations such as interviews or presentations. To that end, it is suggested researchers explore other forms of peer assessment assignment types in Moodle Workshop.

It was mentioned in the limitations that students did not self-assess their own work. This was the intent of the design for this study. However, students could rate their own work as a method of identifying if they are high or low raters. An interdisciplinary perspective could broaden our understanding of the phenomena. Glasser is well-known for his reality therapy theory whereby he believed that individuals consistently rate themselves lower than peers, and many employers use self evaluation as an informative process during job evaluations to identify improvement (Cherryholmes, 1992). Glasser's argument makes sense because people generally know their own weaknesses more than their peers would; thus, self-assessments would serve a useful learning purpose. Moodle Workshop has the capability to allow self-assessments which could be weighted at zero for grading, thereby allowing the comments to serve as reflective constructive feedback to students (formative feedback, but no summative impact on grade).

One suggestion for future research would be for the Moodle software developers to implement a Cronbach alpha reliability or Kappa interrater agreement statistical index into the LMS Workshop module. This suggestion would give the professor a single measure representing the consistency of the student peer assessment activity. It would also indicate which students did not provide a reliable and consistent peer assessment. From that, professors could adjust the student grades and provide constructive feedback to students about their peer assessing skills, substantiated with scientific evidence (rather than observations of the work done). At a minimum it is suggested Moodle software programmers ought to add a grade export feature in Workshop.

Finally, although this study was conducted with a management course in the business and economics discipline, the concepts and practices are readily generalizable to other disciplines. To that end, faculty and researchers are encouraged to investigate this line of analysis with other courses across the disciplines.

It was proposed that peer evaluations might reduce the workload of faculty who are required to teach large sections of classes. Even more importantly, the most strategic benefit is that time of professor would be better spent mentoring students instead of evaluating them. Certainly more studies of this innovative practice will need to be completed and published before we can form any long term practice improvements.

References

Bedore, P., & OSullivan, B. (2011). Addressing instructor ambivalence about peer review and self-assessment. *WPA: Writing Program Administration - Journal of the Council of Writing Program Administrators, 34*(2), 11-36.

Biggs, J. (2003). *Teaching for quality at university: What the student does* (2nd ed.). Buckingham, UK: Society for Research into Higher Education and Open University Press.

Bitter, G. G., & Legacy, J. M. (2008). *Using Technology in the Classroom*. NY: Pearson.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

Cherryholmes, C. H. (1992). Notes on pragmatism and scientific realism. *Educational Researcher*, 14(5), 13-17.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scale disagreement or partial credit. *Psychological Bulletin, 70*(2), 213-220.

Cohen, J. (1992). Statistics a power primer. *Psychology Bulletin, 112*(1), 115-159.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Creswell, J. W. (2009). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (3rd ed.). NY: Sage.

Cronbach, L. J., & Snow, R. E. (1981). *Aptitudes and instructional methods: A handbook for research on interactions* (2nd ed.). New York: Irvington Publishers.

Dollisso, A., & Koundinya, V. (2011). An integrated framework for assessing oral presentations using peer, self, and instructor assessment strategies. *NACTA Journal, 55*(4), 39-44.

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.

Finn, G. M., & Garner, J. (2011). Twelve tips for implementing a successful peer assessment. *Medical Teacher, 33*(6), 443-446.

Gibson, P. A., & Dunning, P. T. (2012). Creating quality online course design through a peer-reviewed assessment. *Journal of Public Affairs Education, 18*(1), 209-228.

Gielen, S., Dochy, F., & Onghena, P. (2011). An inventory of peer assessment diversity. *Assessment & Evaluation in Higher Education, 36*(2), 137-155.

Gielen, S., Dochy, F., Onghena, P., Struyven, K., & Smeets, S. (2011). Goals of peer assessment and their associated quality concepts. *Studies in Higher Education, 36*(6), 719-735.

Gill, J., Johnson, P., & Clark, M. (2010). *Research methods for managers* (4th ed.). London: Sage.

Goda, B. S., & Reynolds, C. (2010). Improving outcome assessment in information technology program accreditation. *Journal of Information Technology Education: Innovations in Practice, 9*(1), 49-59.

Green, S. K., & Johnson, R. L. (2010). Essential aharacteristics of assessment, *Assessment is Essential* (Vol. 6): Mcgraw-Hill.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

Heyman, J. E., & Sailors, J. J. (2011). Peer assessment of class participation: applying peer nomination to overcome rating inflation. *Assessment & Evaluation in Higher Education, 36*(5), 605-618.

Idowu, A. I., & Esere, M. O. (2010). Assessment in Nigerian schools: A counsellors viewpoint. *International Journal of Education Economics and Development, 1*(4), 338-347.

Keppel, G., & Wickens, T. D. (2004). *Design and Analysis: A Researchers Handbook* (4th ed.). Upper Saddle River, NJ USA: Pearson Prentice-Hall.

Li, L. (2011). How do students of diverse achievement levels benefit from peer assessment? *International Journal for the Scholarship of Teaching & Learning, 5*(2), 1-16.

Li, L., & Lei-na, L. (2012). On-line peer assessment of Chinese students oral presentation in English. *Sino-US English Teaching, 9*(3), 1005-1009.

Li, L., Liu, X., & Zhou, Y. (2012). Give and take: A re-analysis of assessor and assessees roles in technology-facilitated peer assessment. *British Journal of Educational Technology, 43*(3), 376-384.

Liu, Z.-F., & Lee, C.-Y. (2013). Using peer feedback to improve learning via online peer assessment. *Turkish Online Journal of Educational Technology, 12*(1), 187-199.

Mok, J. (2011). A case study of students perceptions of peer assessment in Hong Kong. *ELT Journal: English Language Teachers Journal, 65*(3), 230-239.

Mudrak, D. (2011a, January 11). *Best assessment rater scoring in Moodle Workshop 2.0*. www.moodle.org. Available: http://docs.moodle.org/24/en/Using_Workshop [2013, June 1].

Mudrak, D. (2011b, January 6). *Moodle workshop 2.0 specifications*, [Java program]. www.moodle.org. Available: http://docs.moodle.org/dev/Workshop_2.0_specification [2013, June 1].

Ng, E. M. W., & Lai, Y. C. (2012). An Exploratory Study on Using Wiki to Foster Student Teachers Learner-centered Learning and Self and Peer Assessment. *Journal of Information Technology Education: Innovations in Practice, 11*(1), 71-84.

Nulty, D. D. (2011). Peer and self-assessment in the first year of university. *Assessment & Evaluation in Higher Education, 36*(5), 493-507.

Russell, M. K., & Airasian, P. W. (2012). Summative Assessments, *Classroom Assessment. Concepts and Applications* (7th ed., Vol. 5): Mcgraw-Hill.

Sadler, D. R. (2009). Grade integrity and the representation of academic achievement. *Studies in Higher Education, 34*(7), 807-826.

Schroeder, C. M., Scott, T. P., Tolson, H., Huang, T.-Y., & Lee, Y.-H. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching, 44*(10), 1436-1460.

Seery, N., Canty, D., & Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgment in design driven practical education. *International Journal of Technology & Design Education, 22*(2), 205-226.

Shafaei, A., & Nejati, M. (2012). Does student empowerment influence their commitment? *International Journal of Education Economics and Development, 3*(4), 305-313.

Shih, R.-C. (2011). Can Web 2.0 technology assist college students in learning English writing? Integrating Facebook and peer assessment with blended learning. *Australasian Journal of Educational Technology, 27*(5), 829-845.

Speyer, R. e., Pilz, W., Van Der Kruis, J., & Brunings, J. W. (2011). Reliability and validity of student peer assessment in medical education: A systematic review. *Medical Teacher, 33*(11), e572-e585.

Strang, K. D. (2009). Using recursive regression to explore nonlinear relationships and interactions: A tutorial applied to a multicultural education study. *Practical Assessment, Research & Evaluation, 14*(3), 1-13.

Strang, K. D. (2013a, December 1-4). *Exploring summative peer assessment during a hybrid undergraduate supply chain course using Moodle.* Paper presented at the Proceedings of the ASCILITE Electric Dreams Conference, Macqarie University, Sydney, Australia.

Strang, K. D. (2013b). Risk management research design ideologies, strategies, methods and techniques. *International Journal of Risk and Contingency Management, 2*(2), 1-26.

Thomas, G., Martin, D., & Pleasants, K. (2011). Using self- and peer-assessment to enhance students future-learning in higher education. *Journal of University Teaching & Learning Practice, 8*(1), 1-17.

Zhang, A., & Blakey, P. (2012). Peer assessment of soft skills and hard skills. *Journal of Information Technology Education, 11*, 155-168.

Biography

Professor Kenneth Strang has a Doctorate in Project Management (business research, high distinction), an MBA (Honors), a BS (Honors), as well as a Business Technology diploma (Honors). He is a certified Project Management Professional® from Project Management Institute, and he is a Fellow of the Life Management Institute (distinction, specialized in actuary statistics and pension systems), from Life Office Management Association. His research interests include: Leadership, multicultural learning, consumer behavior, and risk management. He designs and teaches multidisciplinary subjects while coordinating the business administation program at the State University of New York (Plattsburgh Queensbury campus) and he also supervises doctoral students. He is the chief editor and associate/area editor of several journals. More information at http://personal.plattsburgh.edu/kstra003/