Exploring summative peer assessment during a hybrid undergraduate supply chain course using Moodle

Kenneth David Strang
School of Business & Economics,
State University of New York, USA;
University of Phoenix; APPC Research, Australia

The main hypothesis of this recent study was that student peer assessing could produce a fair grade in a hybrid undergraduate supply chain course. A key challenge was there were three long written assignments weighted at 90% of the course spread throughout 15 weeks (the final exam in week 16 was weighted at 10%). The secondary goal was to explore if Moodle could facilitate the online assessment of the three project management plans (PPs). A PP was approximately 25 single-spaced pages, based on a unique initiative for each of the 45 students, and it was evaluated against nine Project Management Body of Knowledge (PMBOK) standards as well as other course learning objectives. The PMBOK lectures were classroom-based, data collection was field-based for authentic experiential learning while the LMS was essential for material sharing and assignment management. Interrater reliability, correlation and pair-wise t-test estimates supported the hypotheses. Peer assessments were found to be reliable between students and consistent with the professor's evaluations. Moodle's workshop module was effective but there were two minor shortcomings: (1) reliabilities must be estimated manually, and (2) there was only one rudimentary algorithm in Moodle to calculate the student rater grade for peer assessment quality.

Keywords: summative peer assessment; Moodle workshop; interrater reliability agreement; student peer grading; undergraduate supply chain project management hybrid course.

Introduction

This study contributes to ASCILITE's Electric Dreams community of practice by reviewing the past literature and reporting a successful endeavor which applied summative student peer assessments using Moodle as the Learning Management System (LMS) for a bybrid-mode course (with combinations of classroom, field and online student learning). The findings from this study will inform and motivate future research.

The underlying motive for this study was sadly not merely towards better teaching and learning but rather to improve the efficiency of current practices out of necessity because of increased workload. Faculty must find ways to leverage technology in pedagogy, for continuous improvement and out of necessity due to teaching larger classes. Public university budgets shrank while student enrollment increased in USA as more adults sought degrees to retool or to increase employment competitiveness (United Union Professionals, 2013). Business school faculty are under pressure from disciplinary accreditation entities to increase scholarly research publications and to provide summative assessments which show students are learning (Association to Advance Collegiate Schools of Business, 2013; Accreditation Council for Business Schools and Programs, 2013).

The macro level problem was that many universities in USA had expanded prior to the economic recession by opening additional campuses yet now they found themselves with more students and less budget. In fact, one of the cost-cutting measures taken by the collective group of State University of New York institutions (SUNY) during 2010 was to replace the expensive Blackboard and Angel LMS commercial products with open source

Moodle (developed in Australia). Moodle is free, in terms of purchase price, but it requires expertise from faculty to properly leverage it for pedagogy. Moodle also requires considerable support from the technology staff to implement and manage it. Collectively, SUNY institutions have approximately 468,000 students and over 3 million alumni. Thus, Moodle best practices developed at SUNY are likely to be worth sharing with the global educational community.

The micro level dilemma was that many SUNY institutions needed effective ways to teach large face-to-face courses across different campuses. In the case study, the professor needed to teach the same multi-section face-to-face course at two physical campuses, with a four-hour return drive between them through the Adirondack Mountains in New York State during winter. Many universities around the world face comparable challenges delivering effective education programs to their clients across geographic distances, so those practitioners may be interested in the outcomes of this study.

The course content was predominately qualitative in nature, which required students to produce essays and give periodic presentations of their progress. One of the learning objectives required students to conduct peer assessments, which therefore had to be structured in a way so that fair grades could be given. Students were evaluated on their ability to fairly assess peers (10% of course points), and in turn, students were assessed by their peers (yet the professor was responsible for ultimately evaluating every student outcome).

Peer assessment is an effective pedagogy technique (Falchikov & Goldfinch, 2000; Green & Johnson, 2010; Russell & Airasian, 2012; Speyer, Pilz, Van Der Kruis & Brunings, 2011). Many researchers have published studies demonstrating that peer assessing is helpful to the learning process (Bayat & Naicker, 2012; DiVall et al., 2012; Dollisso & Koundinya, 2011; Finn & Garner, 2011; Gielen, Dochy & Onghena, 2011; Gielen, Dochy, Onghena, Struyven & Smeets, 2011; Heyman & Sailors, 2011; KoÃ, 2011; Kritikos, Woulfe, Sukkar & Saini, 2011; Li, 2011; Li & Lei-na, 2012; Li, Liu & Zhou, 2012; Lu & Law, 2012; Lu & Zhang, 2012; Nicholson, 2011; Nulty, 2011; Shih, 2011; Thomas, Martin & Pleasants, 2011; Wu, Davison & Sheehan, 2012; Wu, Hou & Hwang, 2012; Zhang & Blakey, 2012; Zhi-Feng, Liu & Lee, 2013).

When Falchikov and Goldfinch (2000, p. 314) 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." This indicated that a large portion of these were valid based on the benchmark of 0.70 for reliability (Hair, Black, Babin, Anderson & Tatham, 2006; McCabe, 2007).

There have been various standalone software (e.g., iPeer; WebPA; SPARK) or subsystems (e.g., PeerMark within TurnItIn) for implementing peer assessments. In this study the goal was to utilize Moodle if possible since it was the sanctioned LMS. Moodle has a workshop module specifically developed for facilitating student peer assessments. However, there was no information in the literature as evidence of its effectiveness or about how to configure Moodle for student peer assessments in a face-to-face university business course. More so, there were no guidelines for measuring the consistency of peer assessments. Furthermore, a pilot study by the author using Moodle workshop had revealed that the peer assessment algorithm was not totally reliable, and several experienced Moodle programmers have documented these minor problems (Mudrak, 2011b).

One mandate of this study was to structure pedagogy to include Moodle (technology) to facilitate peer assessment and fair grading of student assignments. Another objective was to statistically measure the interrater reliability of the student peer assessments compared with the professor. An additional goal was to statistically estimate the reliability of using a Moodle workshop for peer assessment. Inductively, recommendations were needed for applying Moodle workshop LMS in higher education practice and for conducting further research.

Literature review

First peer assessment theory in higher education was reviewed, followed by relevant empirical studies. Then the application of peer assessment using LMS technology, specifically Moodle, was researched.

Peer assessment theory

Good quality higher education programs should encourage interaction and use peer assessments in the pedagogy (Johnson & Aragon, 2003). Peer assessments should be used in addition to faculty-generated and self-regulated feedback because students learn best from multiple sources (Strang, 2010b), and through a variety of learning style matches with their professors or tutors (Strang, 2008, 2010a).

Measurement of performance against objective criteria is the fundamental task in a peer assessment, which needs to be clearly structured and simple, in order to be effective for students to administer (Falchikov & Goldfinch, 2000). The words 'assessment' and 'evaluation' are frequently used interchangeably, but they differ in significant ways. Assessments are 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 (Green & Johnson, 2010). Evaluations use the assessments to make judgments about a student's ability and to inform decisions about continued pedagogy (Green & Johnson, 2010). Therefore, peer assessment is concerned with the student grading assignments based on predefined criteria, while faculty will generally evaluate assessment scores to inform ongoing pedagogy.

The words 'formative' and 'summative' are also often mentioned in peer assessments. Formative refers to a pedagogical process done by the professor or students during the course to measure student understanding of the material, as well as to monitor and guide future pedagogy (Russell & Airasian, 2012). Summative 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 assessments are given by the professor as questions posed during the class (or online in a forum) while summative evaluations are done at the end of a learning unit through tests or assignments with predetermined rubrics for grading. Peer student assessments are usually summative in nature (Green & Johnson, 2010) but they could be formative or both depending on the application. "By definition, all student works that contribute to course grades are summative. […] Grades may be pressed into doing double duty: formative and summative" (Sadler, 2009, p. 808). As Sandler implied, formative and summative peer assessments are useful in as far as they provide extrinsic motivation and intrinsic self-efficacy.

The key theoretical problems with peer assessments, including faculty-provided assessments, are reliability, validity, bias and automation with technology. Peer assessment reliability refers to the degree that scores on the assessment are consistent and stable across multiple raters: students, faculty or combinations of both (Green & Johnson, 2010). The three common sources of error in peer assessments which 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 between students and their raters (Green & Johnson, 2010).

A clear design using an objective rubric can reduce bias and improve validity while statistical estimates such as interrater agreement can be generated to measure reliability (Hair, Black, Babin, Anderson & Tatham, 2006; McCabe, 2007; Strang, 2009). A LMS can be used for peer assessments to streamline peer assessment implementation and to improve the effectiveness of the process as well as 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 is obtained by ensuring the rubric is clear. Content-related validity refers to refers measuring the correct objectives. Criterion-related validity refers to using relevant and easy to understand scoring scales, which the raters will use such as nominal, good versus bad wording, or ordinal, e.g., Likert 1 to 10 ratings (Strang, 2009).

Differences between the socio-cultural factors of the rater versus rubric creator versus student often impact the validity and reliability of peer assessments (Li, 2011; Mok, 2011; Shih, 2011). Researchers have argued there will be disagreement between raters regardless of whether they are students or faculty (Falchikov & Goldfinch, 2000). However, the concept behind randomized allocation or peer assessors is derived from the normal distribution in that with enough raters, individual differences should average out (Russell & Airasian, 2012). Evaluator differences also reflect the real world workplace so this is another argument supporting peer assessments.

Falchikov and Goldfinch (2000) acknowledged that faculty may not use peer assessments 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 & O'Sullivan, 2011). Nonetheless, this is an effective learning strategy and pedagogy, in that students learn to improve during the course from the feedback on a formative basis, and faculty may use the assessment scores as part of the grading towards the course learning objectives in a summative manner. Additionally, on the assumption that the student assessing is done fairly, this off loads a large part of the evaluation work from busy faculty when enrollment is large and when the types of assignments are qualitative in nature with long written reports.

Peer assessment studies

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 a pedagogy. They concluded that peer assessment was widely used and it was an effective educational intervention, which improved learning. Their advice for making peer assessment effective was to use an instrument linked to the learning objectives which has high reliability and validity. In effect what they were recommending from empirical experience was to use a rubric to improve objectivity within raters 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 key 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.

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 were generally qualitative assignments which they described as "academic product and process" (Falchikov & Goldfinch, 2000, p. 310), such as reports and presentations.

In their meta-analysis Falchikov and Goldfinch (2000) calculated the correlation R of academic product and process assessments as (0.75 (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 24% which is a large effect (Cohen, 1992). 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 the 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, Xiongyi and Yuchun (2012) conducted a follow up study on this data which 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.

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, but more so later on in the course. An important finding from their work was that students got better at peer assessment with practice. Therefore, an important implication would be requiring students to first complete a practice peer assessment.

Some faculty use peer assessment informally rather than as a grading mechanism. Heyman and Sailors (2011) found that traditional peer assessments helped students learn the material better. They also proposed an interesting approach to better 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.

Peer assessment using technology

One of the well-known scholarly advocates of using technology in education (including peer assessments) was Laurillard (2007). She surveyed 19 higher-education institutions from 13 countries in Asia-Pacific region (Eurpore, Latin America, and North America) to determine the effectiveness of using technology for pedagogy. Her recommendation was to leverage LMS technology to better manage large cohorts.

Bitter and Legacy (2008) emphasized that technology should be subservient to learning objectives when conducting peer assessments through technology. They argued peer assessments are more effective for evaluating (and for student learning) with qualitative assignments, such as team projects and presentations, since there is so much to review, more raters are better able to observe different perspectives to enhance the constructive feedback. Rubrics linked to course learning objective should be designed for objectivity, which refer to competencies, abilities, and attitudes. They offered tips for using peer assessment rubrics in a LMS:

- Avoid highly detailed criteria that become more of a checklist than a rubric;
- Use a limited number of dimensions (aspects, categories);
- Focus on learning priorities of the project;
- Use measurable criteria that can be counted or ranked (such as ordinals, Likert 1-5 or 1-10 scales);
- Use four performance levels that make fine enough discrimination, yet are not too divisive (see below);
- Maintain an equal interval distance between levels so that the highest and next highest are an equal distance to the lowest and next lowest;
- Involve students in creating rubrics so they will clearly understand what the expectations are and this will encourage student support of the process; (adapted from: Bayat & Naicker, 2012; Bitter & Legacy, 2008).

Willey and Gardner (2010) developed a peer assessment model along with a software product called SPARKplus to automate the process. The software could be easily integrated into a LMS. Their model was based on two simple formulas. The first formula 'SPA' was calculated as the square root of total ratings for an individual assessment divided by average of total ratings for al l team members. The limitation for this rating is that it may provide a coefficient larger than 1.0 so a nonlinear correction procedure (manual or programmed) would be needed to implement this for grading purposes. The grade is then calculated by multiplying the SPA by the team score. This cannot be implemented for individual projects as was the case in this study (unless only the SPA coefficient were used with a nonlinear correction algorithm). There were no team projects here only individual projects which were double blind assessed by five other peers. Also their model did not report reliabilities to ensure the peer assessments were consistent, which was essentially the goal of this study.

Thomas, Martin and Pleasants (2011) found the type of technology used for peer assessment did not matter as long as learning objectives were clear. They used wikis for peer assessment at the University of Wollongong. A useful contribution was their recognition that the 'learning value' of peer assessments must be explained to students rather than merely forcing students to use them.

One of the more novel approaches was by Wu, Hou and Hwang (2012) since they used online text messaging as a peer assessment methodology for 38 students. More importantly, they reminded us about the importance of content validity and criterion reliability for peer assessment rubrics. They recommended faculty use Blooms cognitive domain when designing the peer assessment rubric. Interestingly, Lu and Law (2012) published a similar study, echoing the advice to use Blooms Taxonomy to inform the design of the rubric.

Neus (2011) pointed out that raters need to be graded to so as to provide accountability for their peer assessment. However, the biggest issue concerning using peer assessments seems to be how to mathematically calculate grades for the rater (assuming the average of peer ratings would form the score of the rated student). He demonstrated a technique for calculating a correlation coefficient for grading the rater using SAS. The problem with correlation is that since it is a bivariate measure, it works with only two variables, which would mean only up to two raters could be assessed to calculate a 'rater effectiveness' coefficient. In addition, Pearson Product Moment correlation can only be applied to ratio level data not ordinals or intervals such as Likert scales.

Zhang and Blakey (2012) used factor analysis to assign grades to raters for their peer assessments. They were able to validate their rubric assessment scale with Cronbach's reliability values greater than 0.70 and the instrument was able to capture 67% of the variance between rater scores on each assignment. However, factor analysis is a complicated process and it seemed difficult to associate to the rubric.

Dollisso and Koundinya (2011) used paired t-tests to grade raters based on their peer assessments resulting in an effect size of 0.06. The rating scales were 10-point Likert type so these could be considered ordinal data type. Pair wise t-tests would be a labor-intensive technique to assess more than one rater. Nonetheless their concept

has merit since ANOVA is designed to compare the variance of ratings when using ration data while the Kruskall-Wallace test can be used as the nonparametric equivalent of ANOVA when the ratings are in ordinal scales such as the traditional letter grades A-F (Strang, 2009).

Peer assessment in Moodle

The workshop module in Moodle is designed to automate peer assessments. A grade is given for the assessment (from peers) and a separate grade is given to each 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). 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, which function similarly (the latter is more structured).

There is 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 a 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 teacher's 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, 2011a).

The 'best assessment' is determined for each rubric aspect based on finding a 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 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, 2011b).

The grade for assessment (given to a student for assessing peers) is calculated using the 'comparison of assessments' setting in workshop which 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, 2011b). 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%.

Synthesis and research questions

Based on the literature review, peer assessment (automated by a LMS) is a useful to facilitate pedagogy. Peer assessments require a clear rubric without too many criteria items (Bayat & Naicker, 2012; Bitter & Legacy, 2008). In the Moodle workshop module these are called aspects. Likert rating scales from 1 to 10 were recommended (Dollisso & Koundinya, 2011).

The following research questions arose based on the literature review and from the problems noted earlier:

1. Would students rate their peers reliably?
2. Would the student peer ratings be consistent with faculty assessments of the same student assignments?
3. Is the Moodle workshop module effective as a LMS to facilitate student peer assessing?

Methods, procedures and materials

The researcher employed a theory-dependent positivist philosophy consisting of a deductive literature review (above) to inform the research questions, instrument design, and methods (Gill, Johnson & Clark, 2010; Strang, 2013). Since this study was designed to collect performance data, quantitative techniques were selected to

answer the research questions concerning student peer assessment validity and reliability (Creswell, 2009).

Descriptive statistics, correlation, interrater reliability and validity tests were applied at the 95% confidence level. SPSS version 14.1 was used for the statistical tests, while Moodle version 2.4 and workshop version 2.0 were installed at SUNY for this quasi-experiment.

Case study participants

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, with 1050 of those in the School of Business and Economics, of which approximately 350 were in the undergraduate Bachelor of Science in Business Administration (BSBA) program at the time of writing.

At the university level, the average class size was 22, the student-faculty ratio was 17:1, and 97% of tenure-track faculty held the highest degree (e.g., PhD or doctorate) in their discipline. The gender balance was 45.1% male, 54.9% female. International enrollment from 63 countries represented 5.4% of the population.

In the business school 65% of faculty held a relevant doctorate or at least a PhD. The size of this class was 45 due to its demand at both campuses, thus making the ratio 45:1. The researcher had taught large classes of over 600 students so he was familiar with using technology out of necessity to facilitate applying pedagogy in large cohorts. The mean age of the sample was 23 (SD=2.1), while females represented 59% of the class. There were three international students in the sample from different countries (3/45 = 6.7%). The demographic factor and GPA estimates of the sample were similar to the university's business school population (based on z-score tests).

Instrumentation

All 45 participants were undergraduate students in the upper division Project Management (PM) course taught by the researcher. There was one teaching assistant. This course had been taught by the researcher for two years in this context using Moodle, and before that this professor had taught a similar version of this course at other universities using Blackboard, Angel, Moodle and a proprietary LMS. A pilot had been successfully completed in a previous term using an identical course syllabus and with the same configuration in Moodle.

There were four summative assessments, as enumerated below (with course weighting in parenthesis):

- Project management plan 1 (PP1) - knowledge competency or career advancement (20 points);
- Project management plan 2 (PP2) - natural or man-made disaster preparation or mitigation (30 points);
- Project management plan 3 (PP3) - real estate capital investment development (40 points);
- Project management knowledge test - comprehensive and cumulative exam (10 points).

Moodle workshop was utilized for all three PPs. Each PP was around 25 pages. The course weighting for each PP was progressively higher because students were expected to improve their competencies and each PP assignment was more difficult. The format of the PPs were that a multi-page project mandate was presented by the professor then industry subject matter experts were brought in for the students to interview. The grade for each PP was broken into two components: 90% for the charter presentation and plan submission, plus 10% for the quality of the peer assessments performed on other students. The grade for the first component was calculated in Moodle workshop as the un-weighted average of all peer generated scores. The grade for the second component was calculated by Moodle workshop using the 'best assessment' algorithm which was explained earlier.

Students were randomly allocated 5 peer reviewers in Moodle workshop. All peer reviews were based on a rubric (listed in Appendix 1) and each reviewer marks was weighted at 1. The 'comparison of assessments' of fair (2.5) was specified for all PPs. The professor did not complete a review in workshop but instead he manually assessed each PP using the rubric (for experimental control). The ratings were informed by the revised *Taxonomy for Education* (Anderson & Krathwohl, 2001), which range from lowest to highest levels of learning as: remembering, understanding, applying, analyzing, evaluating, and applying (Strang, 2011).

First a mandatory practice PP0 was setup (using a simple class exercise for a General Electric/National Grid project plan) to allow students to become familiar with peer assessing and Moodle workshop. Each PP required students to demonstrate competency in all nine project management knowledge areas. Competencies included using PM software, developing Gantt schedules, applying risk quantification using Program Evaluation and

Review Technique (Strang & Symonds, 2012), and orally presenting the executive summary charter in class through the video conferencing system since two physical campus locations were synchronously linked together for this course. The PM software was OpenProject a free product available from the open software foundation which was similar to Microsoft Project commercial software.

Results, discussion and conclusions

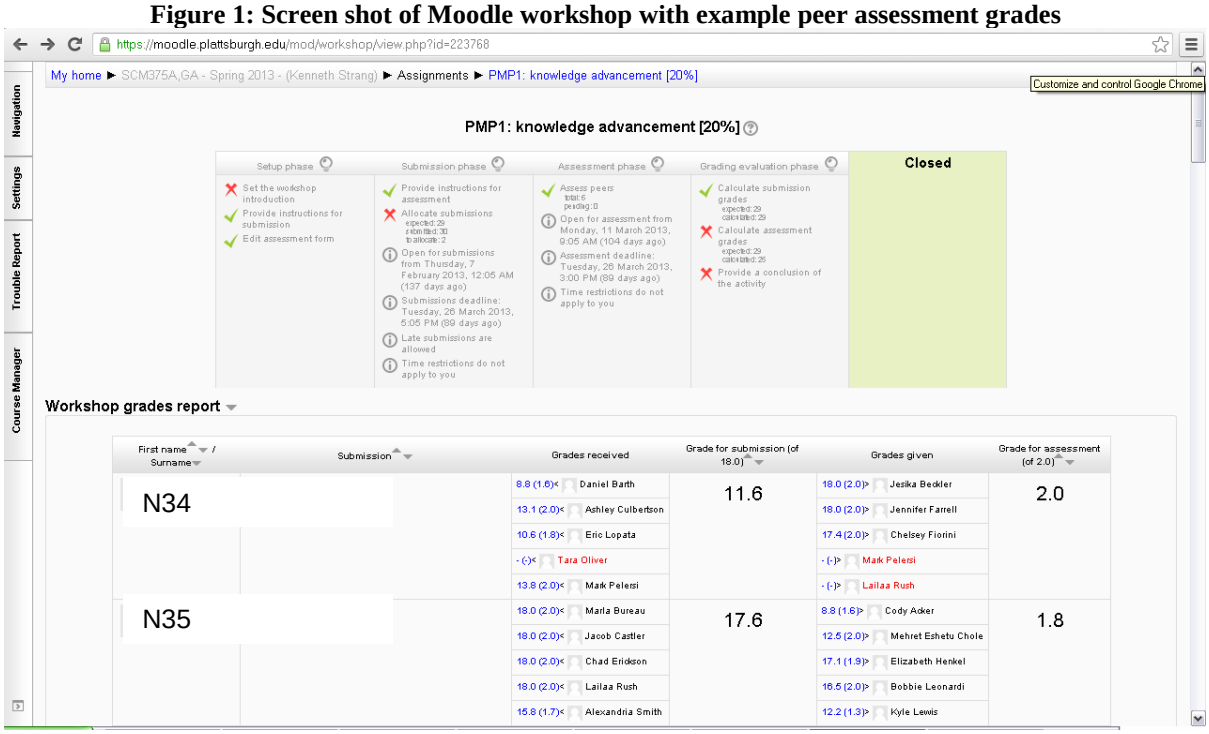
Moodle workshop module implementation

Figure 1 illustrates the first few results for PP1 from the Moodle workshop. The proportion for the submission was set at 18/20 leaving 2/20 for the peer assessment grade. In figure 1, student N34 was given a score of 11.6/20 which Moodle calculated as the mean of peer assessments (multiplied by weights, which were set at 1): (8.8 + 13.1 + 10.6 + 13.8) / 4 = 11.6 (rounded). The score of 2.0 for the peer assessment was calculated based on there being no significant difference between his peer assessment score and the scores from other peers on the same PPs.

N35, the second student in figure 1, received a submission score calculated as the average of: 18 + 18 + 18 + 18 + 15.8 = 17.6 (rounded). His peer grading mark was 1.8/2 (90%) based on the two calculations: difference coefficients = 0.01 + 0.001 + 0.008 + 0.001 + 0.02 = 0.04; peer grading mark = (1 - 2.5 * 0.04/1) = 0.9 * 2 = 1.8.

Based on these results, it appeared that the first research question was supported in that raters were scored reliably using the Moodle workshop 'best assessment' technique. However, additional testing was needed to confirm support for this and to answer the second research question of would the peer ratings be consistent with faculty assessments of the same student assignments.

Figure 1: Screen shot of Moodle workshop with example peer assessment grades



Interrater reliability and comparative reliability tests

In order to answer these questions, interrater reliability was calculated for each PP based on the 5 student raters (or less in a few situations with missing submissions). This can be achieved using a variation of Kappa's interrater reliability (Cohen, 1968) based on the work of Fleiss, Nee and Landis (1979); according to the formula in equation 1.

In equation 1, ***f*** is the Fleiss-Kappa interrater coefficient (higher values mean more consistency), where k = number of Likert scale levels, n = number of rubric aspect categories receiving a k rating, r = number of raters, x^2 = chi square of

$$f = 1 - \frac{nr^2 - \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}^2}{nr(r-1)\sum_{j=1}^{k} p_j q_j} \quad (1)$$

difference between expected and observed ratings for each *k* rating by each *r* rater, p_j = mean proportion for *k* rating *j*, and q_j = compliment of p_j (1 – mean proportion for mean proportion for each *k* rating *j*). Subscripts *i* and *j* are matrix indexes, which point to individual Likert ratings by each rater (*r*) for each rubric aspect (*n*).

The ***f*** was calculated for each student across all 3 PP assignments, whereby all coefficients were above 0.60 and most were close to 0.80. A benchmark for good interrater agreement is generally 0.80 (Cohen, Cohen, West & Aiken, 2003) but some researchers have accepted 0.70 (Hair, Black, Babin, Anderson & Tatham, 2006; McCabe, 2007) which is the benchmark applied in this study.

For example, an ***f*** coefficient was calculated using the data for student N35 as shown in table 1 (scores were scaled to 18 for the PP2 assignment and rounded). PP2 was weighted at 20 points (out of 100 for the course), and the submission plus presentation component was weighted at 90%. Therefore, 90% * 20 = 18 points, leaving 10% * 20 = 2 points for the quality of peer assessment grade. Note that the k value was 5 because the rating scale was zero to 4. The ***f*** kappa (r=5, n=9, k=5) = 0.79, s^2= 0.0075, z =9.138, p=0.000, N=35 (DF=31), with control intervals for the ***f*** (0.62, 0.96). This 79% coefficient was a statistically significant result with acceptable interrater agreement based on research practices (Hair, Black, Babin, Anderson & Tatham, 2006) - however this was illustrated for the peer assessments on only a single student PP. The kappa's were manually calculated for all assessments with a mean of 0.80 interval (0.62, 0.96). Thus there was preliminary support of the first research question that students gave their peers a fair grade.

Table 1: Peer assessment grades for student N35 (M=17.6; 5 raters, 9 aspects, 5 scale levels)

Rubric criterion (aspect)	Marla	Jacob	Chad	Lailaa	Alex
Integration management	4	4	4	4	4
Scope management	4	4	4	4	4
Time management	4	4	4	4	4
Cost management	4	4	4	4	4
Risk management	4	4	4	4	4
Human resource management	4	4	4	4	4
Quality management	4	4	4	4	4
Communications management	4	4	4	4	2
Procurement management	4	4	4	4	2
Score (scaled to 18 total)	18	18	18	18	16

Then to answer the second research question, the average peer-generated score for each PP was compared with the professors' score, which was estimated statistically by applying a paired t-test. A two-tailed test was selected because the goal was to test the inequality of the Moodle workshop grade (average of peer ratings) as compared to the grade given by the professor. The results of the paired t-test supported the research question, D (134) = 0.31, p=0.76 (two-tailed). In this case, it was desirable to see no significant difference between scores.

The nonparametric Spearman correlation was very high between each students ***f*** and the peer assessment grade given in Moodle (score out of 2): Rho r = 0.92, p=0.000 (two sided), n=45 students, N=135 assignments. Spearman correlation is more conservative than Pearson Product Moment and the former does not assume a normal distribution underlies the evaluation results (furthermore we cannot expect students to grade on the curve or that there ought to be 68% of the mean ratings with one SD of the mean for a peer assessment). Therefore, the third research question was accepted in that Moodle was useful in managing the peer assessing process and the algorithm calculated a fair peer grade to each student which was similar to the kappa (92% correlation). This was proven by comparing the workshop peer grade score for all 135 assignments to the ***f*** interrater agreement.

Limitations and recommendations

A key limitation in this research, which affects any generalizations, was the small sample size of 45 students. Additional the context of SUNY may not be similar to other universities. For example the international composition of this SUNY institution was 5.4% (from 63 countries) and there were three international students in the sample (3/45 = 6.7%). Furthermore, this quasi-experiment was applied on business school undergraduate students. Fourthly, the professor's pedagogical approaches may differ substantially from others.

Notwithstanding the above, there was strong evidence to support both research propositions that Moodle workshop can be effectively used for peer assessments. There were no instances of students receiving an

incorrect peer grade and the high correlation of 92% between peer grades and Fleiss-Kappa interrater reliability for each assignment indicated a high level of consistency. However, this study should be replicated with larger samples, across other disciplines, at different institutions, in different socio-cultures, and in online modality.

The researcher did not locate any other LMS, which provided a peer assessment module as Moodle did. This is also an area of recommended future research - to provide peer assessment modules for the other LMS products.

Implications and future research

Moodle workshop was effective. There was statistical support in that there was no significant difference between the professor grading versus the student peer assessments on all 3 assignments (N=45 students).

One suggestion for future research would be for the Moodle developers to implement a Kappa statistical score into workshop, which could provide another peer grading alternative. Furthermore, it would provide faculty with statistical estimates of how well the students were performing regarding their peer assessments. 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).

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 4.7 out of 5 for the instructional items on the survey (SD=0.6, N=37 respondents).

In closing, 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 still assessed every student assignment in this course (N=135), if the Kappa interrater reliability statistic had been available, he could have just randomly sampled a few, thus saving 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.

For example, the researcher took on average 20 minutes to assess each project plan in this course. There were 3 * 45 = 135 project plans (excepting that one student did not submit a PP1 due to illness). Therefore, assuming other professors would take similar time to assess such assignments in other courses, a total of 270 minutes would be needed for this activity. If the professor instead merely sampled 10% of the assignments, based upon the potential availability of a built-in Kappa interrater reliability statistic (or having access to SPSS to calculate this), and further assuming the students were capable of assessing peer assignments (as was this cohort), the professor would save 270 * 90% = 243 minutes or about 4 hours every course. If this savings were extrapolated across the entire school of business at this university for a year, it was estimated that the time equivalent to another faculty position would be saved. Imagine the potential benefits if this concept of technology-facilitated student peer assessing were applied at all business schools and in other disciplines? This might be an effective pedagogy if a reliability coefficient was calculated and reported in the LMS Moodle workshop module.

References

Association to Advance Collegiate Schools of Business (AACSB). (2013). *Eligibility Procedures and Standards for Business Accreditation*. Tampa, FL: AACSB. Retrieved from http://www.aacsb.edu/accreditation/process/documents

Accreditation Council for Business Schools and Programs (ACBSP). (2013). *Preliminary Visit Questionnaiure for Baccalaureate / Graduate Degree Schools and Programs*. Kansas City, KS: ACBSP. Retrieved from http://www.acbsp.edu

Anderson, L., & Krathwohl, D. (2001). *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York, NY USA: Longman

Bayat, A., & Naicker, V. (2012). Towards a learner-centred approach: interactive online peer assessment. *South African Journal of Higher Education, 26*(5), 891-907

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

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

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*, 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. ISBN: 9781412965569
DiVall, M., Barr, J., Gonyeau, M., Matthews, S. J., Amburgh, J. V., Qualters, D., et al. (2012). Follow-up assessment of a faculty peer observation and evaluation program. *American Journal of Pharmaceutical Education, 76*(4), 1-7
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
Fleiss, J. L., Nee, J. C. M., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin, 86*(1), 974-977
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. ISBN: 978-1847870933
Green, S. K., & Johnson, R. L. (Eds.). (2010). *Essential Characteristics of Assessment* (Vol. 6): NY: 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
Johnson, S. D., & Aragon, S. R. (2003). An instructional strategy framework for on-line learning environments. *New Directions for Adult and Continuing Education, 10*, 31-44
KoÃ, C. (2011). The views of prospective class teachers about peer assessment in teaching practice. *Educational Sciences: Theory & Practice, 11*(4), 1979-1989
Kritikos, V. S., Woulfe, J., Sukkar, M. B., & Saini, B. (2011). Intergroup peer assessment in problem-based learning tutorials for undergraduate pharmacy students. *American Journal of Pharmaceutical Education, 75*(4), 1-12
Laurillard, D. (2007). Modelling benefits-oriented costs for technology enhanced learning. *Higher Education, 54*(1), 21-39
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 assessee's roles in technology-facilitated peer assessment. *British Journal of Educational Technology, 43*(3), 376-384
Lu, J., & Law, N. (2012). Online peer assessment: effects of cognitive and affective feedback. *Instructional Science, 40*(2), 257-275
Lu, J., & Zhang, Z. (2012). Understanding the effectiveness of online peer assessment: A path model. *Journal of Educational Computing Research, 46*(2), 313-333
McCabe, M. (2007). Assessment, blending and creativity: The abc of technology in mathematics teaching. *The International Journal for Technology in Mathematics Education, 14*(1), 54-59
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. Retrieved June 1, 2013, from http://docs.moodle.org/24/en/Using_Workshop
Mudrak, D. (2011b, January 6). Moodle Workshop 2.0 Specifications. Retrieved June 1, 2013, from http://docs.moodle.org/dev/Workshop_2.0_specification
Neus, J. L. (2011). Peer assessment accounting for student agreement. *Assessment & Evaluation in Higher Education, 36*(3), 301-314

Nicholson, D. T. (2011). Embedding research in a field-based module through peer review and assessment for learning. *Journal of Geography in Higher Education*, 35(4), 529-549.
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. (Eds.). (2012). *Summative Assessments* (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
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. (2008). Quantitative online student profiling to forecast academic outcome from learning styles using dendrogram decision models. *Multicultural Education & Technology Journal, 2*(4), 215-244. Retrieved from http://dx.doi.org/10.1108/17504970810911043
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. Retrieved from http://pareonline.net/getvn.asp?v=14&n=3
Strang, K. D. (2010a). Global culture, learning style and outcome: an interdisciplinary empirical study of international students. *Journal of Intercultural Education, 21*(6), 519-533. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/14675986.2010.533034#preview
Strang, K. D. (2010b). Measuring self-regulated e-feedback, study approach and academic outcome of multicultural university students. *International Journal of Continuing Engineering Education and Life-Long Learning, 20*(2), 239-255. Retrieved from http://www.inderscience.com/search/index.php?action=record&rec_id=36818
Strang, K. D. (2011). A grounded theory study of cellular phone new product development. *International Journal of Internet and Enterprise Management, 7*(4), 366-387. Retrieved from http://www.inderscience.com/browse/index.php?journalID=39&year=2011&vol=7&issue=4
Strang, K. D. (2013). Risk management research design ideologies, strategies, methods and techniques. *International Journal of Risk and Contingency Management, 2*(2), 1-26. Retrieved from http://www.igi-global.com/article/risk-management-research-design-ideologies/77903
Strang, K. D., & R. J. Symonds (2012). Analyzing research activity duration and uncertainty in business doctorate degrees. *International Journal of Risk and Contingency Management*, 1(1), 29-48.
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
United Union Professionals (UUP). (2013). *Suny Budget and Student Enrollment Trends*. Albany, NY: UUP
Willey, K., & Gardner, A. (2010). Investigating the capacity of self and peer assessment activities to engage students and promote learning. *European Journal of Engineering Education*, 35(4), 429-443.
Wu, K., Davison, L., & Sheehan, A. H. (2012). Pharmacy students' perceptions of and attitudes towards peer assessment within a drug literature evaluation course. *American Journal of Pharmaceutical Education, 76*(4), 1-4
Wu, S.-Y., Hou, H.-T., & Hwang, W.-Y. (2012). Exploring students' cognitive dimensions and behavioral patterns during a synchronous peer assessment discussion activity using instant messaging. *Asia-Pacific Education Researcher, 21*(3), 442-453
Zhang, A., & Blakey, P. (2012). Peer assessment of soft skills and hard skills. *Journal of Information Technology Education, 11*, 155-168
Zhi-Feng Liu, E., & Lee, C.-Y. (2013). Using peer feedback to improve learning via online peer assessment. *Turkish Online Journal of Educational Technology, 12*(1), 187-199

Appendix 1: Peer assessment rubric applied in Moodle workshop

Aspect (category)	Criteria	Rating (0-4)
Integration management	a. first page has project correct title, PM name, date, course number; b. project is unique (and approved online in PMIS); c. version log is present and realistically completed (version 1 or similar); d. table of contents if accurate and well formatted; e. charter mentions key items from scope such as key deliverable and reason for project, overall cost and time, PM; f. all other eight sections are present; g. APA references at end for citations to sources; h. spelling, grammar, and professional business writing and speaking evident at all times; i. uploaded in PDF format with OpenProj or Planner Gantt file attached.	
Scope management	a. indicated exact nature of project; b. some background (with a citation to literature or news article); c. start and complete date (or duration); d. at least one key deliverable (relates to reason for doing project); e. at least one assumption; f. at least one constraint; g. spelling, grammar, and professional business writing and speaking evident at all times.	
Time management	a. includes Gantt with tasks shown; b. at least 3 resources (PM + 2); c. sequence and links can be seen; d. indentation used with WBS numbers; e. at least one milestone visible; f. formatted clearly and professionally with no duplication from risk or other sections; g. spelling, grammar, and professional business writing and speaking evident at all times (including timely delivery of charter briefing presentation).	
Cost management	a. includes external and internal unit costs summarized by category; b. at least 2 levels of detail (categories); c. overall total; d. earned value formula shown; e. earned value calculation correct; f. SPI and CPI shown as percentages; g. implications on budget discussed; h. no duplication from procurement or other sections; i. spelling, grammar, and professional business writing and speaking.	
Human resource management	a. at least 3 resource roles explained; b. costs shown (including PM); c. listed in table format (resource allocation matrix); d. unit costs given; e. same resources as shown on Gantt chart; g. material resources shown; f. person resources used; g. spelling, grammar, and professional business writing and speaking.	
Risk management	Internal risks: a. method for estimating internal risks listed; b. table included (well formatted, labeled, referenced in text), shows risky tasks, and overall critical path method risk (standard deviation); c. shows probability project will finish 10% earlier than expected duration.; External risks: d. method for estimating external risks listed; e. identification (2-3 likely risks listed applicable if the project were underway); f. sources for risks noted (subject expert interviews); g. spelling, grammar, professional business writing and speaking.	
Quality management	a. at least 3 key (reasonable) quality goals identified; b. key criteria in a matrix (table with heading etc), c. selection of method to measure quality explained; d. formulae or benchmarks identified; e. citations to quality guidelines; f. zero defects; h. spelling, grammar, professional business writing and speaking.	
Communication management	a. at least 3 key (reasonable) stakeholders identified; b. key deliverables in a matrix (table with heading etc), c. emails for stakeholders to notify them; d. mention use of technology or method for above; interviews); e. communication matrix complete with W5+how format; g. spelling, grammar, professional business writing and speaking during project charter briefing with other PM's.	
Procurement management	a. includes external and internal unit costs for materials summarized by category; b. at least 2 levels of detail (categories); c. overall total; d. contract types explained; e. justification for contract types given; f. implications on budget discussed; h. no duplication from cost or other sections; i. spelling, grammar, and professional business writing and speaking evident at all times.	
Ratings are scores of competency or proficiency, informed by the revised *Taxonomy for Education*, where: 0 is the lowest and 4 is the highest: 0 = not addressed, 1 = basic understanding but many requirements missing and typos, 2 = application of key requirements but typos and some items missing, 3 = sound analysis but typos or a few requirements missing; 4 strong demonstration of knowledge area with all requirements met.		

Author details

Dr Kenneth David Strang, Doctorate, MBA, BS, BT, FLMI, CNA, PMP
School of Business and Economics
State University of New York
640 Bay Road, Regional Higher Education Building, Queensbury, NY, USA 12804
Tel: +1 518 792 5425
Fax: +1 518 792 3868
Web: http://personal.plattsburgh.edu/kstra003/

Please cite as: Strang K.D. (2013). Exploring summative peer assessment during a hybrid undergraduate supply chain course using Moodle. In H. Carter, M. Gosper and J. Hedberg (Eds.), *Electric Dreams. Proceedings ascilite 2013 Sydney*. (pp.840-853)