Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

A Measure to Cultivate Engaged Peer Assessors: A Validation Study on its Efficacy

Yu-Hsin LIU[a], Kristine LIU[b] & Fu-Yun YU[c*]
[a]*Department of Civil Engineering, National Chi Nan University, Taiwan*
[b]*Medill School of Journalism, Media, Integrated Marketing Communications, Northwestern University, USA*
[c]*Institute of Education, National Cheng Kung University, Taiwan*
**fuyun.ncku@gmail.com*

Abstract: Despite the generally positive learning effects of peer assessment, undesirable behaviors exhibited during the process have been reported (e.g., peer assessors engaging at a superficial level, or giving biased judgements). With reference to related literature and based on non-participant observation of student assessors' behavior in classrooms and document analysis of past student assessors' work, a measure consisting of four variables was devised to serve two purposes: on the passive side, to alleviate such reported hindrances; on the active side, to engage peer assessors in sensible, prudent ratings. The devised measure quantifies the performance of student assessors based on the scores they give to their peers' performance/work as compared to that of the teacher/expert on each assessment criterion and two other noteworthy variables (i.e., fine discrimination ability, completion rate). A validation study which involved two classes of university sophomores was conducted. The students presented individual projects while participating in assessing their peers' performance via an online system, which only differ in whether the devised measure was in use (the experimental group, $N = 53$) or not (the comparison group, $N = 47$) for the two respective class. The statistically different results in peer assessors' performance between the two treatment groups, $t\ (98) = 8.97 < .001$, attested the efficacy of the devised measure in encouraging higher quality peer assessments.

Keywords: Expert ratings, performance assessment, peer ratings, online peer assessment, the quality of peer assessment

1. Introduction

Peer assessment has been a subject of investigation since Topping's (1988) highly cited review paper. Studies investigating the potential of peer assessment for the support of the teaching and learning process have mushroomed over the years. Empirical studies accumulated over the last three decades have generally confirmed the positive effects of peer assessment for promoting academic performance (Double, McGrane, & Hopfenbeck, 2020; Li, Xiong, Hunter, & Guo, 2019), deeper learning (Li, Bialo, Xiong, & Hunter, 2020; Sluijsmans et al., 2002), core 21st century skills development (Yu & Wu, 2016), and positive social-affective outcomes (vanGennip, Segers, & Tillema, 2009).

In light of the many affordances of networked technology, online peer assessment has been an emerging area that continues to attract increasing attention from academics and educators (Kulkarni et al. 2015). Currently, many online peer assessment platforms are available in the market to support the assessment of students' produced work/performance and the effort and time invested by the involved collaborative group-members. While supportive evidence on learning gains from these developed systems has been reported, some undesirable signs and behaviors exhibited during peer assessment have been noted. Examples of these aspects include peer assessors engaging at a superficial level and giving biased, unfair judgements (Adachi, Tai & Dawson, 2018; Liu & Carless, 2006; Yu & Sung, 2019). Hence, devising efficacious designs to help alleviate these challenges of online peer assessment is a worthwhile and important endeavor.

In this work, a peer assessment measure is proposed. The peer assessment measure, in the form of an equation, quantifies the quality of the assessors' assessment and consists of four variables. A

validation study was conducted to provide preliminary data on the efficacy of the devised peer assessment measure.

2. An Online Peer Assessment Measure to Promote Engaged Process

Knowing that the relative accuracy of peer and teacher/expert ratings is a major concern of educators (Li et al., 2015), the correlation between peer and teacher/expert ratings takes center stage in the devised equation. In addition, non-participant observation of student assessors' behavior exhibited during the peer assessment process in classrooms and document analysis of past student assessors' ratings of their peer work/performance from the previous academic year were employed to identify other noteworthy variables — differentiation of peer work/performance along each criterion and completion rate. In short, the equation devised to quantify the quality of each student assessor's performance consists of four variables: $Y = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4$, where w denotes the respective weighting associated with each variable. Each of the four *x* variables is explained below.

Considering that peer assessment criteria may be subjective (e.g., organization, logic, clarity, appeal, interest, visual design, conciseness, etc.) and objective (e.g. within the limit of time or page) by nature, variable x_1 deals with all criteria of a subjective nature whereas variable x_2 deals with the objective type. Both variables x_1 and x_2 are obtained by calculating the correlation between a student assessor's score with that of the teacher/expert on each respective criterion. With reference to Li et al. (2015) meta-analysis which found the estimated average Pearson correlation (i.e., *r*) between peer and teacher ratings to be moderately strong (i.e., $r = .63$ to be exact), the calculated *r* is translated to a score by referring to Tables 1 and 2 for subjective and objective criteria, respectively. Generally, the student assessor and teacher/expert's low *r* corresponds to a low score (with calculated *r* below 0 being given a scoring of 0) whereas higher *r* corresponds to a higher score.

Compared to subjective criteria (x_1), objective criteria (x_2) is expected to have a clear-cut benchmark to compare against. Thus, it would demand a higher correlation between the score of the student assessor and that of the teacher/expert than subjective criteria. As shown in Tables 1 and 2, when the calculated *r* equals to or higher than 0.83, it is given a full score of 100 for subjective criteria. As for objective criteria, the calculated *r* would need to equal to or higher than 0.90 to be given a full score of 100. Also, the translated scores of all criterion are summed and averaged for subjective and objective criteria, respectively.

Table 1. x_1 *Scores Corresponding to Given r Values for Subjective Criteria*

Calculated *r* value	Scoring
[0.83, 1.00]	100
(0.63, 0.83)	100*(0.17 + *r*)
0.63	80
[0.00, 0.63]	100*(0.17 + *r*)
(0.0, -1.0)	0

Table 2. x_2 *Scores Corresponding to Given r Values for Objective Criteria*

Calculated *r* value	Scoring
[0.90, 1.00]	100
[0.00, 0.90]	100*(0.10 + *r*)
(0.0, -1.0)	0

For the x_3 component, we adopt the concept of item discrimination as stressed by psychometrics to denote the ability of a test item to discriminate amongst examinees (Amedahe & Asamoah-Gyimah, 2016). The spread and range of student assessors' scoring on each assessment criteria are considered in our devised equation. To this aim, standard deviation (denotes *sd*), which connotes the spread of scores one gives to their peers, is used to see if student assessors possess the intended fine differentiation on each criterion. Conceptually speaking, high *sd* would represent high differentiation ability while concentration of scoring around a rating scale is characterized by low *sd*. Nonetheless, to account for undesired possible polarized scoring, which would also result in high *sd*, x_3 is proposed to be calculated

as listed in Table 3. Here, M = maximum possible value of *sd*, which will change depending on the number of points in a scale. M is calculated as: (the number of points in a scale - 1)/2. For example, for a 7-point scale, M = (7-1)/2 = 3.

Table 3. x_3 *Scores Corresponding to Calculated sd Value*

Calculated *sd* value	Scoring
[0.00, (1/3)M)	100*(3/M)**sd*
[(1/3)M, (2/3)M]	100
((2/3)M, M]	100*(3/M)(M – *sd*)

The last variable attributing to the quality of student assessors' performance, x_4, considers peer assessment task completion. It is calculated by counting the total number of peer assessment forms individual student assessor submitted against the expected number of forms to be submitted. The score the student assessor receives is proportional to the rate of which s/he completes the assigned peer assessment. Student assessors with a 100% completion rate are granted a full score on this metric.

3. A Validation Study on the Proposed Measure Efficacy

3.1 The Participants, Context, and Online Learning System

Two classes of college sophomores (N = 100) taking a required three-credit course (i.e., Transportation Engineering) from a university located in central Taiwan participated in this validation study. As part of the course requirement, the participants were asked to make an oral presentation with PowerPoint on their chosen topic. They were also asked to assess their classmates' presentation. The participants were informed that both their performance on PowerPoint presentation and peer assessment activity accounted for 20% of their final grade.

An online system to support online peer assessment activity developed by the authors was extended by embedding the devised equation to be activated by the instructor. Each of the participants would rate their peers' presentation on each of the devised criteria of the instructor's choice on their choice of personal devices.

3.2 Research Design, Data Analysis, and Test Results

To test the efficacy of the devised equation, a validation study involving two intact classes was conducted. Both participating classes used the same online system for the peer assessment activity. The only difference between the two group lies in whether the equation function is activated (the experimental group, N = 53) or not activated (the comparison group, N = 47). The participants of the experimental group were simply told that with reference related literature, a measure to objectively quantify their performance at peer assessment activity was in place. However, the participants had no knowledge as to what variables were involved or how their assessment scores were calculated.

For this validation study, five criteria were devised for the peer assessment activity. They are the overall visual design of PowerPoint slides, oral communication (e.g., fluency, clarity, logic), content (e.g., organization), control of presentation time, and presence and appearance. The participants were directed to rate their peers' presentation on a 7-point Likert scale on each of the five criteria. Minimal instruction was provided for the four subjective criteria as the participants' judgements may vary contingent on individual liking. Nonetheless, for objective assessment, a scoring instruction was developed for the objective criterion (i.e., the 'control of presentation time' criterion of this study) (see Figure 1 for the score to be given for the assessment of control of presentation time.

Figure 1. Scoring instruction given and posted for objective criteria — control of presentation time

In this validation study, the instructor of this course played the role of the expert. The expert respective scores given to individual presentation was compared to that given by each student assessor in a criterion-by-criterion fashion and then averaged before being translating to x_1. For the validation test, the weighting for the four variables were arbitrarily set at: 0.4, 0.2, 0.2, and 0.2 for x_1, x_2, x_3, and x_4, respectively, where each of the four subjective criteria accounts for 10% of presentation score.

Descriptive statistics of the two participating classes in peer assessment performance are listed in Table 4. Further independent sample *t*-test performed found a statistical difference between the two treatment groups, t (98) = 8.97 < .001. In other words, integrating the devised measure in the online peer assessment system helps promote higher quality peer assessment.

Table 4. *Descriptive Statistics of Peer Assessment Performance in the Two Treatment Groups*

Treatments Statistics	The comparison Group (N = 47)	The experimental Group*(N = 53)
Mean	62.74	84.98
Standard Deviation	15.94	6.27

*The devised measure integrated in the adopted online peer assessment system

4. Conclusions

The learning effects associated with peer assessment have generally been confirmed positively. Yet, a couple of undesirable signs and behaviors exhibited during online peer assessment have been noted. With reference to related literature and based on non-participant observation and document analysis of student assessors' behavior and ratings in the previous semester, a measure consisting of four variables was devised to engage peer assessors in prudent ratings while alleviating implicit personal bias from affecting peer evaluation scores. As evidence by the validation study, the devised measure is efficacious in promoting overall better peer assessment performance.

References

Adachi, C., Tai, J. H-M, & Dawson, P. (2018). Academics' perceptions of the benefits and challenges of self and peer assessment in higher education, *Assessment & Evaluation in Higher Education, 43*(2), 294-306.

Amedahe, F. K., & Asamoah-Gyimah, K. (2016). *Introduction to measurement and evaluation* (7th ed.). Cape Coast: Hampton Press.

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review, 32*, 481–509.

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D. &, S. R. Klemmer (2015). Peer and Self Assessment in Massive Online Classes. In H. Plattner, C. Meinel, and L. Leifer (eds) *Design thinking research* (pp. 131–168). Cham: Springer

Li, H., Bialo, J., Xiong, Y., & Hunter, C. V. (2020). Effects of peer assessment on students' non-cognitive outcomes. *American Educational Research Association Annual Meeting*. San Francisco.

Li, H., Xiong, Y., Hunter, C. V., & Guo, X. (2019). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education 45*(1),1-19.

Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., K., Lyu, Y., Chung, K. S., & Suen, H. K. (2015). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245-264.

Liu, N.-F. & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education, 11*(3), 279–290.

Sluijsmans, D. M. A., Brand-Gruwel, S., van Merriënboer, J. J. G., & Bastiaens, T. J. (2002). The training of peer assessment skills to promote the development of reflection skills in teacher education. *Studies in Educational Evaluation, 29*(1), 23–42.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249-276.

van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educational Research Review, 4*(1), 41–54.

Yu, F. Y. & Sung, H. S. (2019). Online targeting behavior of peer-assessors under identity-revealed, created, and concealed modes. *Educational Technology and Society, 22*(1), 15-27

Yu, F. Y. & Wu, C. P. (2016). Predictive effects of the quality of online peer-feedback provided and received on primary school students' quality of question-generation. *Educational Technology & Society, 19*(3), 234-246.