Asian Journal of the Scholarship of Teaching and Learning Volume 8 ■ Number 1 ■ June 2018

Article

Improving Peer Grading Accuracy in MOOCs: The Effects of Prior Experience, Time Spent on Grading, and Training

Nikan Sadehvandi

Centre for the Promotion of Excellence in Higher Education, Kyoto University, Japan

Correspondence:

Name: Nikan Sadehvandi
Phone: (+81) 080-8522-3662
Address:No. 402, Konoekan, Bldg 69, Yoshida-Konoe Street, Sakyo-ward, Kyoto, 606-8315 Japan.

Email: sadehvandi.nikan.5n@kyoto-u.ac.jp

Recommended citation:

Sadehvandi, N. (2018). Improving peer grading accuracy in MOOCs: The effects of prior experience, time spent on grading, and training. *Asian Journal of the Scholarship of Teaching and Learning, 8*(1), 25-47.

ABSTRACT

This paper reports on a study conducted on a course offering of KyotoUx to investigate the effect of training on improving the accuracy of peer assessment. In addition, this study explores the effects of two other variables: the time spent on grading and the prior peer assessment experience on students' grading accuracy. The results of the analyses indicated that inter-rater agreement was higher between those peer graders who had higher levels of involvement with the training materials. Also, there were significantly higher correlation indices between peer graders more involved with the training materials and the teaching assistant with respect to their grades on the same submission. The time spent on grading and prior assessment experience did not turn out to be major influencing factors on students' accuracy of grading in this study. We made several suggestions to further optimise the effectiveness of peer assessment in MOOCs and put forth several directions for further research.

INTRODUCTION

As enrolment numbers grow in MOOCs, the instructor's grading of open-ended assignments in such courses has become more impracticable. Thus far, MOOCs have relied on different assessment techniques to gauge learners' grasp of the content being covered throughout a given course. For example, machine-graded multiple-choice quizzes and problem sets have been applied widely to assess students' knowledge retention of the subject at hand (Kulkarni *et al*., 2013). Recently, attempts have also been made to automate the scoring of learners' literary products (Balfour, 2013). However, even in its most sophisticated form, automated essay algorithms can only capture shallow aspects of a written work such as grammatical errors, words count, or text length.

Currently, peer grading is among the few viable methods that can be deployed on a massive scale and at the same time, take into account meaningful features of students' writing such as analytical reasoning and creativity, which is otherwise difficult, if not impossible, to capture through automated grading. However, due to the large number of students participating in MOOCs, an instructor cannot supervise the peer grading activity and regrade an assignment if needed. Therefore, addressing questions such as to what extent are peer assessment results for MOOCs trustworthy, and whether these results reflect students' unbiased and adept judgment on a given work, has become an emerging issue in MOOC literature (Suen, 2014; Staubitz, Petrick, Bauer, Renz, & Meinel, 2016). In addition to monitoring for learner-related factors that might influence the consistency of peer grading, previous research has reported the need for training in order to increase the accuracy of peer grading performance on different aspects of a grading task (Reddy & Andrade, 2010).

The last few years has witnessed a trend towards designing and implementing different models to prompt students to do a better assessment. For example, Kulkarni and colleagues (2013) conducted a controlled experiment to improve learners' grading through feedback. They compared learners' grading and staff grading on the same assignment and based off of that, learners were informed whether their grading was close to that of the staff's or the extent to which their grading deviated from it—too high or too low. Staubitz, Petrick, Bauer, Renz, & Meinel (2016) employed a feedback rating approach in which learners rated the review of peers from a scale of one to three based on how good they thought the quality of a given review was. This was done using a guideline on what constitutes a good review and the reason was to allow students to make an impact on their grades and to defend their work, which in turn incentivised peer reviewers to generate high quality assessments and earn them the chance of receiving at most one bonus score per review they would write. While the number of such studies on improving the quality of peer assessment in MOOCs is gradually increasing, they are still few in number, mostly conducted in a

single course, and their findings can only be generalised to similar subjects and a similar population of learners. Therefore, there is a need to extend this line of research on improving the quality of peer assessment in MOOCs with different subjects and populations of learners. This paper reports on a study which investigates the effects of training on one form of peer assessment, namely, peer grading, through the use of a support mechanism and two other factors—time spent on grading and prior experience—on accuracy of peers' awarded grades in one of the course offerings at KyotoUx[1] .

REVIEW OF THE RELATED LITERATURE

Most MOOCs rely on peer assessment as it offers logistical benefits where the number of submitted works exceeds the capacity of a human grader. In other words, while human grading is not scalable in a MOOC, peer assessment makes grading of tens of thousands of students' works possible. However, the application of peer assessment in MOOCs also poses some challenges. According to Suen (2014), MOOC participants reportedly question the quality of peer assessment, resulting in their ultimate aversion to participating in such activities and sometimes quitting the course altogether. Suen, citing Jordan (2013) also maintains that attrition rates tend to be considerably high in MOOCs that implement peer assessment, although for him it is not clear whether this is a side effect of peer assessment per se or due to the open-ended format of the assignment tasks learners need to submit.

Among a myriad of rater-related factors associated with grading that could influence both the process and product of assessment, the raters' experience appears to be a frequently researched factor (Choonkyong, 2016). However, the reported findings were mixed, ranging in result from those in which the type of rating scale affected the way raters used their experience in rating (Barkaoui, 2011) to those in which time was a major factor in bringing inexperienced and experienced raters closer in grading (Joe, Harmes, & Hickerson, 2011). Although the need for investigating the role of prior experience on MOOC peer assessment is being gradually recognised (Yousef, Chatti, Wosnitza, & Schoroeder, 2015), there is still a paucity of research to address this factor within a MOOC setting.

Another factor that has captured the interest of peer assessment researchers is the amount of time learners spend on grading activities as a function of learners' efforts towards accurate grading. Piech *et al.* (2013) maintain that spending an adequate amount of time—what they refer to as "a sweet-spot"

[1] This refers to Kyoto University's platform on edX.

created by normalised grading time—on peer review is associated with more accurate grades, whereas spending excessive or too little time could signal a flawed evaluation on the part of the peer reviewer. They further argue that spending less time than required is associated with a less meticulous peer review process resulting in grade inflation, while spending more time than necessary may indicate that the reviewer might have experienced some predicaments in doing the assessment (e.g. they may have difficulty understanding some aspects of the scoring rubric).

Research suggests that attempts towards systematising peer assessment, such as introducing a scoring rubric, could ensure more consistent grading (Jonsson & Svingby, 2007). Yet, mere use of a scoring rubric does not always ensure that the awarded grade reflects the true quality of the task being assessed. This is because different raters interpret and apply the grading criteria in different ways. It has been argued that upon discussing, negotiating, and implementing the assessment criteria by all learners, both the "assesse" and assessor would become more cognisant of the key elements of assessment and what might constitute high quality work (Topping, 1998). In the same fashion, the existing evidence suggests that raters must be trained in accordance to the scoring rubric (Reddy & Andrade, 2010). Moreover, an ideal training on peer assessment should comprise: 1) sample responses on different performance levels accompanied by adequate justification (from an evaluation authority) to explain the reasons behind the accorded grade, and 2) the possibilities for rating practice with opportunities to compare one's grading with the instructor's on the same task (Meier, Rich, & Cady, 2006). The question to address would then be this: To what extent would such training (of specific guidelines and procedures in assessment) fulfill the attempt to increase the reliability and validity of peer assessment results?

The purpose of this study was twofold. First, the study empirically explored the effects of different levels of learner involvement, with a support mechanism on peer assessment designed by the researcher, on raising the accuracy (i.e. consistency and validity) of peer assessment results and its perceived effect on learner's satisfaction with the awarded grades. Additionally, the study investigated the role of two other learner attributes: learners' prior experience with peer assessment in MOOCs, and the time spent on peers' grading contributions. The purpose was to find out how the two aforementioned rater-related factors influenced learners' grading behaviour and the variation in learners' awarded grades.

METHOD

Context of the study

The study was conducted for a single homework assignment in the KyotoUx course "002x: Culture of Services: New Perspectives On Customer Relations (First Design)". The course was conducted over eight weeks and focused on the social and cultural aspects of services. The course content incorporated video data of authentic cases in which services were offered to customers (e.g. sushi bars, restaurants, apparel stores, etc.). The homework assignment, named "Homework Assignment 3-3", required learners to analyse a video on the interaction between an employee and a customer in a casual apparel store and write their answers to several homework questions based on the video's content. The scores for "Homework 3-3" counted towards 12% of the final grade. Learners were asked to assess their peers' homework assignments using a grading rubric, which will be explained later in this paper. A minimum of two grading contributions was required of each learner. The number of learners who submitted this homework was 341, of which only 160 proceeded with the peer grading activity.

Course peer grading rubric[2]

The researcher initially speculated that learners would do a more accurate job in peer grading if they had more information on what would constitute the right answer from the instructor's perspective. This information was provided by the course instructor in the form of an ideal response to the same homework question that learners needed to grade; the purpose was to create a certain level of standardisation by bringing peer grading more in line with the instructor's. To this end, a rubric was created which contained two questions on whether the submission captured the "important points" and "additional points" as stated in the instructor's sample response. For this purpose, the rubric also included the instructor's sample response for each homework question which learners could refer to when assessing their peers' submissions (Table 1). Learners were required to grade the submissions, looking for instances of important points and additional points in their peer responses.

---

[2]. The generic term "rubric" is adopted by most open online platforms to refer to a simplified set of criteria to be used during the peer assessment process. It differs in structure from a typical rubric since it does not include any descriptor of the level of performance in a given criteria.

Table 1

Provision of instructor's sample responses to each homework question

Homework questions		Instructor's sample responses	
1.	What does the employee do prior to this talk? Describe how she moves and observes the customer.	Important points	After working on the clothes across what the customer was looking at, she takes on clothes and steps to the left of the customer, about three meters apart. After arriving to the left of the customer, the employee looks at the customer and then returns to the original location.
		Additional points	While stepping to the customer's left, she says, "Welcome," not to the customer but to nobody. This talk to nobody can be heard by the customer that the employee was talking to another customer than herself and therefore moved to her left.
2.	What does the customer do prior to this talk? Describe her gaze.	Important points	The customer holds up a piece of clothes and then looks right, far and then left. These actions show that she is clearly looking for something, most likely a mirror or possibly a fitting room.
		Additional points	Note that she does not necessarily look at the employee who was right in front of her.
3.	When and how does the employee speak to the customer? Describe her move in reaction to the customer's move.	Important points	The employee observes these actions of the customer and then quickly returns the clothes she was holding to the original place, and walks toward the customer to guide her to the mirror.
		Additional points	As she approaches the customer, the customer directs her gaze from right to the employee. The employee is ready to guide the customer to a mirror.

They also needed to assess these instances across three scales of "Good", "Satisfactory", and "Needs Improvement". For each of the three homework questions, the scale "Good" was defined as responses that contained all the important⁄additional points, "Satisfactory" for responses that only partially contained the inclusion of important⁄additional points, and "Needs Improvement" for responses which did not cover any important⁄additional points (see Table 2 for a detailed description of the rubric).

Table 2

Course peer grading rubric

Dimensions \ Scales	Good	Satisfactory	Needs Improvement
1. Does your peer's response to this question capture the important points in the instructor's sample response?	Completely	To some extent	No
2. Does your peer's response to this question capture the additional points in the instructor's sample response?	Completely	To some extent	No

Support mechanisms for improving peer grading

Peer grading (PA) training

It is technically possible for the course design and development team to embed a training component into the peer assessment module on the edX platform. Participation in the peer assessment training can be mandatory or optional, depending on the way it is defined by the course development team. The peer assessment training also provides a good avenue for the course instructor to bring learners' grading closer to their own by means of "calibration" (Balfour, 2013). In this step, learners need to grade a few sample submissions that have already been reviewed and graded by the instructor. After some attempts, the grades that the learners give to these sample submissions should ideally be approximate to those granted by the instructor before they proceed to grade their peers' submissions.

In Week 4 of the course, the training was activated for the learners and participation was optional, that is, the points awarded for participating in the training were not included in the final evaluation. The reason for this was to minimise the learners' perceived burden of doing extra work, which could potentially lead to more dropouts during the peer assessment activity.

The PA training comprised three following steps:

1. **Submit your response:** In this step, learners had to answer a question about an interaction video between a customer and an employee in an apparel store, and submit their responses for peer grading. This was similar to what learners were expected to do when completing the homework.
2. **Learn how to assess responses:** After submitting their responses in the previous step, learners could practice doing peer assessment, grading a peer's submitted answer, using a maximum of two attempts. Next, learners

could view the instructor's grading of the same submitted answer at the end. If the learners' grading and the instructor's grading matched in the first attempt, they could proceed with the peer assessment activity for Homework 3-3. If not, learners' could make another attempt by grading a different submission.

3. **Assess peers**: This was where the peer graders were given guidance to perform the peer grading activity on submissions for Homework 3-3. As previously mentioned, learners needed to grade at least two submissions by the designated deadline.

Rubric-aligned instructional video

The researcher created a five-minute instructional video with voiceover narration, which was embedded in the peer grading activity for Homework 3-3. The aim of the video was twofold: to enable learners to navigate through the training steps easily, and use the rubric to the fullest during the peer grading activity.

The instructional video included the following elements:

1. A description of the purpose of the peer assessment
2. A step-by-step guide on the training steps explained before
3. An extensive elaboration of the grading rubric and its three scales (i.e. good, satisfactory, needs improvement)
4. A set of exemplars, which reflect "good", "satisfactory" or "needs improvement" responses for both the essential and additional points
5. A reminder for the start date and due date of the peer grading activity.

The instructional video was an optional add-on. Nonetheless, learners were reminded and encouraged to watch it before conducting the peer grading activity.

Datasets and instruments

The data sources used in this study included an anonymised record of the learners' submission IDs, "learners as scorers" IDs, "learners as scorees" IDs, their grades, as well as the submission records for Homework 3-3. These were entered into an excel spreadsheet and grouped in the initial screening process based on students' levels of involvement with different features of the support mechanism. In addition, the researchers also captured and analysed data relating to learners' interaction with the instructional video.

The researchers also devised several survey questions, which were embedded into different sections of the courseware. These questions, accessible via a dropdown menu, asked learners about their prior experience with MOOC peer assessment and the average amount of time spent on peer grading. Learners were also asked about their perceptions on whether their peers were graded fairly and their level of satisfaction, choosing from seven response options across a Likert scale, ranging from "1–Strongly Disagree" to "7–Strongly Agree".

Quasi-experimental design and setup

In the initial screening of the data, the IDs of peer graders whose grades did not reflect the true quality of the works assessed were excluded from the analysis. This was done qualitatively, by scanning the content of submissions which were given either a very low or a very high grade. We also looked into the number of peer graders who evaluated a single submission and the number of submissions which were graded by a single peer graders to potentially include in our analysis. However, as the retrieved IDs for these peer graders and submissions were very few in number, the analysis only included the IDs of peer graders that were associated with grading two submissions and were graded by two peer graders in return. The researcher also excluded peer graders' IDs that had accessed the support features after the submission deadline for the peer grading of Homework 3-3 had passed. As a result, the data was reduced to 180 sets of two peer graders, each set grading the same submission.

Based on the learners' type and levels of involvement with different features of the support mechanism, three experimental conditions were set:

1. **Peer graders who fully received the treatment**. These refer to peer graders who took the PA training and interacted with the instructional video, i.e. the "full-treatment" condition.
2. **Peer graders who partially received the treatment.** These refer to peer graders who interacted with the instructional video but did not take the PA training, i.e. the "partial-treatment" condition.
3. **Peer graders who did not receive the treatment**. These refer to peer graders who did not use any of the support elements provided, i.e the "control" condition.

The researcher could have listed another experimental condition, which included peer graders who only took the training step. However, as there was only one learner who took the training step without interacting with the instructional video, this number was deemed too trivial to form the third experimental group. Moreover, because the distribution of submissions to

learners on edX was done through the MOOC's default peer review system, prior selection of subjects into different treatment conditions was not an option. As an alternative, the researchers observed possible distribution patterns of the two peer graders of a single homework submission based on the three aforementioned experimental conditions. This observation yielded five patterns which were called "Case Types" for the ease of reference.

As shown in Table 3, Case Type #1 includes the same assignments evaluated by a peer grader from the control condition and one from the partial-treatment condition. Case Type #2 involves assignments graded by a peer grader from the control condition and one from the full-treatment condition. In Case Type #3, a peer grader from the full-treatment condition and one from the partial-treatment condition evaluated the same assignments. In Case Types #4 and #5, both peer graders belonged to the partial-treatment condition and full-treatment condition, respectively.

Table 3

Assignment of two peer graders to grading submissions based on the three experimental conditions (n=180)

Case Type	*Control*	*Full-treatment*	*Partial treatment*	*Number of observations*
#1	PG_1		PG_2	12
#2	PG_1	PG_2		26
#3		PG_1	PG_2	59
#4			PG_1, PG_2	24
#5		PG_1, PG_2		59

Note: PG=Peer Grader

RESULTS

Reliability and validity of peer grading results across five group types

Intraclass correlation coefficient (ICC) (Shrout & Fliess, 1979) one-way random effects were conducted to calculate inter-rater agreement between pairs of grading on the same homework in each of the five aforementioned case types, separately. According to Shrout & Fliess (1979), this model is appropriate when data consists of different raters who are selected from a bigger pool of raters, which was also the case of the peer graders in this study. In the section that follows, we report both single and average measures of ICC. The former is the reliability of a single peer grader while the latter has to do with the average score of multiple peer graders.

As presented in Table 4, the coefficient values drawn from the analysis of both single and average measures ICCs are for Case Type #3. It indicates that although the coefficient value of .28 for single measure ICC is of a small strength (Cohen, 1988, pp. 79-81), a single peer grader of Case Type #3 could provide more reliable grades than peer graders of other case types. With respect to average measure ICC for Case Type #3, the coefficient value of .44 indicates a moderate strength, which means that the average score of the two peer graders is even more reliable than a single peer grader. Next, the highest single and average measure ICCs were for Case Types #5 and #4, respectively. As one might expect, both single and average measure ICCs for Case Types #1 and #2 were the lowest of all.

Table 4

Intraclass correlation between peer graders across different case types (n=180)

Case Types	*#1 (n=12)*	*#2 (n=26)*	*#3 (n=59)*	*#4 (n=24)*	*#5 (n=59)*
Single Measures	.082	.078	.287	.215	.225
Average Measures	.151	.145	.446	.354	.367

Although not statistically significant, the inter-reliability coefficients were still higher for those case types with deeper levels of involvement with the support mechanism (i.e. either with the instructional video, training step, or both features).

Another proxy used to establish peer-grading accuracy was to find out how closely did the peer graders' assigned grades match with those given by the teaching assistant on the same homework submission. For this purpose, the teaching assistant regraded the same submissions for Homework 3-3 that the peer graders had evaluated before. Table 5 summarises the result of the bivariate correlation analysis between mean-based grades awarded by the two raters in each case type and the grades awarded by the teacher assistant for the same submission.

Table 5

Correlations of peers' mean-based grades and teacher assistant's grades

Case Type	*Measure*	*M*	*SD*	*r*	*p*
1 (*n*=12)	Peers	14.66	6.44	.28	.164
	Teacher-assistant	11.58	5.14		
2 (*n*=26)	Peers	16.25	5.70	.60	.371
	Teacher-assistant	16.19	5.64		
3 (*n*=59)	Peers	18.01	5.43	.66**	.000
	Teacher-assistant	17.23	4.75		
4 (*n*=24)	Peers	18.20	4.80	.57**	.003
	Teacher-assistant	16.70	5.83		
5 (*n*=59)	Peers	17.03	4.77	.54**	.000
	Teacher-assistant	16.17	5.16		
Overall (*n*=180)	Peers	16.77	5.09	.54**	.000
	Teacher-assistant	16.76	5.53		

*p < 0.05, **p < 0.01.

According to the results shown, the teaching assistant's mean score was less than peers' mean score for each case type. However, the peers' (M=16.77) and teaching assistant's grades (M=16.76) were almost the same for all the assessments. The mean-based grades by both peers' and teaching assistant's grades significantly and positively match for those case types in which both raters were either fully or partially engaged with the support features. As can be seen in Table 4 in the previous section, among the five case types, the highest significant correlation was for Case Type #3 (r=.66, p<0.01), which was deemed to be large in strength (Cohen, 1988). The correlation between mean-based grades from Case Type #4 (r=.57, p<0.01) and Case Type #5 (r=.54, p<0.01) were also quite similar in strength, although slightly lower than that of Case Type #3. The correlation between overall peers' mean-based grades and the teaching assistants' grades was also significant (r=.54, p<0.01).

Interrelationship between time spent on peer grading, prior experience, and peer grading

Table 6 presents the frequency analysis of learners' reported time spent on grading their peer submissions for Homework 3-3. Given that each learner had to grade at least two submissions, the reported time referred to the average amount of time the learners spent grading each of the two submissions. The

course instructor proposed that a five- to ten-minute interval was sufficient for learners to conduct a proper assessment of the submissions.

As one can observe in Table 6, the majority of peer graders in the "control" condition (77%), the "partial treatment" condition (84%), and the "full treatment" condition (74%) either selected the target amount of time or the adjacent options. Also, among the three experimental conditions, 22% of the peer graders in the "control" condition, 9% in the "partial treatment" condition, and 21% in the "full treatment" condition reported that they spent longer amounts of time (i.e. 20~30 mins, 30~60 mins, > 60 mins) to grade the submissions.

Table 6

Learners' reported amount of time on peer grading for each experimental condition

Experimental conditions	*N*	*< 2 min*	*2 to 5 min*	***5 to 10 min***	*10 to 20 min*	*20 to 30 min*	*30 to 60 min*	*> 60 min*	*Mean*
Control	9	0%	**22%**	**33%**	**22%**	0%	11%	11%	3.7
Partial-treatment	33	3%	**27%**	**24%**	**33%**	3%	9%	0%	3.3
Full-treatment	65	3%	**26%**	**30%**	**18%**	12%	6%	3%	3.4

Note: "5 to 10 min" is the instructor proposed time interval needed to grade Homework 3-3

Regarding the number of MOOC peer assessment activities that peer graders had taken prior to enrollment in this course, 67% of raters in the "control" condition (*M*=1.7) reported that they had either taken zero (56%) or only one (11%) MOOC peer assessment activity. In the "partial treatment" condition (*M*=3.1), 44% reported that they either had no prior experience of assessing other peers in a MOOC (27%) or they had done one (10%) or two (17%) activities prior to this course. Moreover, 64% of the peer graders in the "full treatment" condition (*M*=2.3) reported that they were either unfamiliar (43%) or slightly experienced with MOOC peer assessment (Table 7).

Table 7

Learners' reported number of peer assessment activities in prior MOOC offerings for each experimental condition

Experimental conditions	*N*	*Zero*	*One*	*Two*	*Three*	*Four*	*Five*	*Six or more*	*Mean*
Control	9	56%	11%	0%	11%	0%	0%	22%	1.7
Partial-treatment	30	27%	10%	17%	3%	0%	3%	40%	3.1
Full-treatment	61	43%	10%	11%	3%	3%	2%	28%	2.3

Note: Zero=Inexperienced; One or Two=Slightly Experienced; Three or Four=Somewhat Experienced; Five or Six=Experienced

Meanwhile, Pearson correlation was conducted to measure the relationship between learners' average amount of reported time spent on grading Homework 3-3 submissions, their two awarded grades on different submissions, and their prior experience with MOOC peer assessment for the three experimental conditions. As presented in Table 8, there was no significant correlation among the four aforementioned covariates. The correlation indices obtained between the average time spent on peer grading and prior experience with MOOC peer assessment were of little value, indicating that there was no statistically significant correlation among the two covariates (i.e. average amount of time spent on peer grading and the level of prior experience with peer assessment in MOOCs) and peers' grades for Homework 3-3.

Table 8

Correlation between average time spent on grading, grading contributions, and prior experience with MOOC peer assessment

Experimental conditions		*r*	*p*		*r*	*p*
Full	AATSPG	-.23	.061	AATSPG	-.15	.203
	FG			SG		
	LPEPA	.08	.516	LPEPA	.05	.673
	FG			SG		
	AATSPG	-.01	.907	AATSPG	-.05	.907
	LPEPA			LPEPA		
Partial	AATSPG	-.08	.637	AATSPG	-.02	.900
	FG			SG		
	LPEPA	.13	.447	LPEPA	.29	.102
	FG			SG		
	AATSPG	-.01	.914	AATSPG	-.01	.914
	LPEPA			LPEPA		
Control	AATSPG	-.07	.642	AATSPG	.02	.888
	FG			SG		
	LPEPA	.06	.711	LPEPA	.19	.222
	FG			SG		
	AATSPG	-.05	.729	AATSPG	-.05	.729
	LPEPA			LPEPA		

FG–First grading, *SG*–Second grading, *AATSPG*–Average amount of time spent on peer grading, *LPEPA*–Level of prior experience with peer assessment

Standard multiple regression analysis was also conducted separately on each set of grading. This was done to establish how much variance in all assessments could be explained by the two measures, namely, the average amount of time spent on grading and the level of prior experience with peer assessment in MOOCs. Table 9 summarises the descriptive statistics for the four variables:

Table 9

Descriptive statistics for total number of peer graders

	Mean	*Std. Deviation*	*N*
First grading	15.8704	7.48468	108
Second grading	17.7130	6.59375	108
Average time spent	3.4444	1.39648	108
Prior experience	3.3241	2.53141	108

The two sets of grading tend to be different [*M*(first grading)=15.87, *M*(second grading)=17.71], suggesting that there exists variance in the way peer graders have have been treating different assignments. Also, peer graders spent approximately the expected amount of time on grading their peers (*M*=3.44) and they were somewhat experienced in conducting peer assessment activity in the MOOC (*M*=3.32).

The multiple squared correlation coefficient for the first and second sets of grading were .032 and .016, indicating that the independent measures accounted for 3.2% and 1.6% in the first and second sets of grading respectively. Table 10 shows that the two control variables were not significant predictors of the learners' first [F(2,105)=1.761, $p>0.05$] and second sets of gradings [F(2,105)=.832, $p>0.05$]. Neither of the two measures were found to significantly predict variations in the dependent variable. Moreover, neither of the levels of prior experience with peer assessment in MOOCs [*t*(first grading)= .828, $p>0.05$, *t*(second grading) =.899, $p>0.05$] and the average time spent on grading [*t*(first grading)= -1.66, $p>0.05$, *t*(second grading) =-.904, $p>0.05$] were found to be significant predictors of the learners' two sets of awarded grades.

Table 10

Standard multiple regression analysis in predicting the two sets of grading by a single peer

	R	R^2	R^2_{adj}	*F(2,105)*	*P*	β	*t*	*P*
FG	.180	.032	.014	1.761	.177			
AATSPG						-.160	-1.665	.099
LPEPA						.080	.828	.409
SG	.125	.016	-.003	.832	.438			
AATSPG						-.088	-.904	.368
LPEPA						.087	.899	.371

FG–First grading, *SG*–Second grading, *AATSPG*–Average amount of time spent on peer grading, *LPEPA*–Level of prior experience with peer assessment

Learners' perceptions of fairness of and satisfaction with the peer grading results

The learners' feedback and opinions were collected through a post-treatment survey. Two follow-up survey questions where designed and embedded into the courseware asking learners—that is, those who received the assessment as opposed to those who did it—about their perceived fairness of and satisfaction with the grades. Learners' IDs were checked with their corresponding peer graders' IDs for the two survey questions in order to divide these learners based on the case type that their peer graders belonged to. Table 11 summarises all the learners' responses on the two survey items.

First of all, it appears that all learners, irrespective of the case type their peer graders belonged in, perceived their grades as being fair and were satisfied with the grades they received. Interestingly, there is an increasing trend in the positive proportion of responses from Case Type #1 through to Case Type #5 for both survey questions, indicating that students who received their grades from peers with higher levels of involvement with the support mechanism tended to be more content with the peer grading.

Table 11

Learners' ratings of two survey items on perceived fairness and satisfaction with awarded grades across five case types

Survey items from "Strongly Disagree" to "Strongly Agree" (1-7)*	*Scored by case type*	*N =98*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*Positive*	*Mean*
I believe that all the peer assessors graded my Homework 3-3 fairly and reasonably.	#1	8	0	0	0	3	2	0	3	62%	5.3
	#2	20	0	2	2	4	2	3	7	60%	5.1
	#3	28	0	1	2	4	2	5	13	64%	5.5
	#4	14	0	0	1	1	1	7	4	85%	5.8
	#5	28	1	1	0	4	5	6	11	78%	5.6
I am satisfied with the grades that I received from my peers for my Homework 3-3.	#1	8	0	0	0	4	1	0	3	50%	5.2
	#2	20	2	2	1	5	2	2	6	50%	4.6
	#3	28	0	0	2	7	1	4	14	67%	5.7
	#4	14	0	0	1	3	0	6	4	71%	5.3
	#5	28	1	0	1	3	4	8	11	82%	5.7

*"Slightly Agree", "Moderately Agree", and "Strongly Agree" were considered as positive responses in this analysis.

DISCUSSION

The ICC analysis of raters' grading contributions in each case type yielded fair and modest indices, albeit statistically non-significant. However, the result of bivariate correlation between the same raters' grades and the teaching assistant's ranged from significantly moderate to high indices at a significance level of 0.01. As the research suggests, the inter-rater inconsistency is an inevitable source of error and this can never be completely eliminated (Wang, 2010). Yet, training can reduce the systematic error to a certain degree as was indicated by the results of this study. The results also corroborate previous findings on the necessity to use multiple raters to ensure high consistency (Cho, Schunn, & Wilson, 2006). In this study, a minimum of two peer graders were required to complete peer grading on a single grading task, and this number was deemed to yield the lowest consistency index (Luo, Robinson, & Park, 2014). Case Type #1 grades had the greatest mean difference with that of the teaching assistant, indicating that these peer graders tended to grade the submissions more generously and had more variance to their grading than the teaching assistant. This could be attributed to differences in the grading abilities of the two peer graders as the result of the treatment (i.e. "control" condition vs. "full treatment" condition), which also explains the lower agreement index established for this case type. Moreover, it has been pointed out that if the assessment criteria is not clear, learners' assessment poorly

correlates with the instructor's assessment (Avery, 2014). In this study, those peer graders with higher levels of involvement with the support mechanism (i.e. "partial-treatment" condition and "full-treatment" condition) supposedly had higher chances of comprehending the course grading rubric than those who did not. Thus, these peer graders' assessment correlated higher with the teaching assistant's on the same submission. All these findings were in favour of the support mechanism, denoting that even an expedited period of training can have perceivable effects on raising the inter-rater agreement and validity of the grading results.

Contrary to the premise put forth by Piech *et al*. (2013)—that time spent to review a work is a factor influencing how well a learner grades—in this study, time was not a determining factor associated with learners' grading accuracy. Meanwhile, the findings corroborate with Alfaro and Shavlovsky's (2016) findings on the relationship between time spent on peer review and grading accuracy. In their study, Alfaro and Shavlovsky (2016) obtained a weak correlation that did not indicate error peaks in learners' grades as a result of spending little time on peer assessment. They attributed the short duration of time spent by learners on reviewing their peers' works to either the learners' ability to perform a quick evaluation of their peers' work or their unwillingness to take the peer reviewing process seriously.

As mentioned earlier in this paper, the diversity of learners' prior experience was highlighted as a factor influencing the effectiveness of peer assessment in MOOCs (Yousef, Chatti, Wosnitza, & Schroeder, 2015b). In a study on the roles of rating scales and rating experience, and their interaction on rating English as Second Language (ESL) essays, Barkaoui (2011) found that the rating experience did not have a major influence on participants' rating processes. In a study on the effect of a short traning programme for evaluating responses to an essay writing task, Attali (2015) reported that the newly trained raters' ability in scoring approximated to that of the experienced group of raters after the training. With regard to prior experience of learners with peer assessment, this study did not find any significant results in favor of peer graders with higher levels of experience. However, as indicated by the analysis of the pearson correlation, the "partial treatment" group tended to slightly inflate the grading contributions, even though the average peer assessment activity taken by this group in previous MOOCs turned out to be slightly greater than the others. On the one hand, experienced raters have been reported to have paid less consistent attention to rubric features and attended to those aspects that were not listed on the grading rubric (Joe , Harmes, & Hickerson, 2011). Also, it might be the case that different grading rubrics evoke rater experiences in different ways. Thus, the grading experience is a complex entity that should be considered in light of interaction with other grading-irrelevant factors depending on the context and guidelines of the assessment. Also, another explanation could be

that the "partial treatment" group might have felt overconfident and performed a cursory assessment of their peers' works due to their familiarity with the type of peer assessment that was conducted in the MOOC, as reported by these learners. Finally, the findings from this study also corroborate the conclusions reached by Luo, Robinson, and Park (2014) on the need for a minimum of three peer graders to evaluate a single submission to ensure higher levels of inter-rater agreement and validity.

CONCLUSION AND FURTHER RESEARCH

The work presented in this paper aimed to study the effects of three variables—training, time spent, and prior experience—on improving the accuracy of peer grading in MOOCs, and gave evidence of the potential benefits of the support mechanism. Overall, as indicated by the results of this study, the training had positive effects on improving peer graders' agreement. Peer graders with higher levels of involvement with the support mechanism tended to have higher levels of agreement in their grading. The training also helped peer graders to give grades that were more similar to that of grades given by the teaching assistant. However, prior experience and time spent on grading did not turn out to be major factors in determining the accuracy of peer grading.

The support mechanism provided to facilitate peer grading was well-received by the learners of 002x. Almost 90% of peer graders of the homework in question were somehow—either fully or partially—engaged with the support mechanism. This may indicate that most learners valued the support mechanism. Despite the potential unforeseen problems in the design of such instructional features, such support mechanisms could still be promising avenues to secure more reliable and valid peer grading outcomes and participants' satisfaction with evaluation results thereof.

It should be kept in mind, however, that both the process and product of a grading task could be influenced by many other interacting factors that can always introduce bias into the grading process. Although peer grader's bias can never be thoroughly eliminated, training is one way to mitigate the influence of such factors. Perhaps, collecting and maintaining learners' profiles on the level of experience and grading proficiency, content knowledge, research interest during the initial stages of the course could benefit the instructor or teaching assistant in terms of helping them to more accurately match up peers of the same grading ability later on. In this study, low inter-rater agreement between the two raters in cases with lower levels of involvement with the support mechanism might not have been entirely due to the lack of training. We believe that the degree of peer graders' compliance with the grading rubric

might differ from one grader to another. This might occur because learners are weak in interpreting aspects of the rubric or are simply lazy to pay close attention to its details, which might have been the case in this study. Thus, we believe that care should be taken in generalising the results and we strongly argue that it is of prime importance to think of ways to incentivize the optimum application of the grading rubric by the learners.

Moreover, we could have conducted an interview with peer graders to shed more light on other factors affecting learners' bias towards inflating/deflating their grading contributions (i.e. the "partial treatment" group). However, few peer graders were willing to cooperate with the researcher in this regard. We believe that more research is still needed to determine the learner-related factors that might affect the accuracy of peer grading. Lastly, the support mechanism was administered without supervision from the researcher. More efforts would be required to exert control over the peer grading and the training process to monitor learners' depth of engagement with the training material, e.g. the length of time spent watching the instructional materials, or how attentive learners were during the training phase. Finally, the experiment reported in this paper spanned over the course of a single peer assessment activity. Therefore, the researcher could not investigate the changes which occurred prior to and after the intervention. Future research may be needed to repeat this study in order to strengthen the findings.

ACKNOWLEDGEMENTS

The authors would like to thank Associate Professor Yutaka Yamauchi and the KyotoUx production team for their wholehearted support, which made this research possible.

ABOUT THE AUTHOR

Nikan Sadehvandi is a Ph.D. candidate at the Centre for the Promotion of Excellence in Higher Education at Kyoto University. She also works as a research associate in the same centre. Her teaching and research interests include, but are not limited to, open education, active learning, and peer assessment.

REFERENCES

Alfaro, L. D. & Shavlovsky, M. (2016). Dynamics of peer grading: An empirical study. *Proceedings of the 9th International Conference on Educational Data Mining*, 62-69. Retrieved from http://www.educationaldatamining.org/EDM2016/proceedings/paper_23.pdf.

Attali, Y. (2016). A comparison of newly-trained & experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99-115. http://dx.doi.org/10.1177/0265532215582283

Avery, J. (2014). Leveraging crowdsourced peer-to-peer assessments to enhance the case method of learning. *Journal for Advancement of Marketing Education, 22*(1), 1-15. Retrieved from http://www.mmaglobal.org/publications/JAME/JAME-Issues/JAME-2014-Vol22-Issue1/JAME-2014-Vol22-Issue1-Avery-ppl-15.pdf.

Balfour, S. P. (2013). Assessing writing in MOOCs: Automated essay scoring and calibrated peer review. *Research & Practice in Assessment, 8*. Retrieved from https://pdfs.semanticscholar.org/2c03/81a33a503179eaa74630576cc7be5dc23eb4.pdf.

Barkaoui, K. (2011). Think-aloud protocols in research on essay writing: An empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51-75. http://dx.doi.org/10.1177/0265532210376379

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*(4), 891-901. http://dx.doi.org/10.1037/0022-0663.98.4.891

Choonkyong, K. (2016). Raters training for scoring rubrics: Rater-centered bottom-up approach. *Minne TESOl Journal*. Retrieved from http://minnetesoljournal.org/wp-content/uploads/2015/12/ChoonKimRater-centered11-17-15.pdf.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Joe, J. N., Harmes, C., & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring: A mixed methods application to oral communication assessment. *Assessment in Education: Principles, Policy & Practice, 18*(3), 239-258. http://dx.doi.org/10.1080/0969594X.2011.577408

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144. http://dx.doi.org/10.1016/j.edurev.2007.05.002

Jordan, K. (2013). MOOC completion rates: The data. Retrieved from http://www.katyjordan.com/MOOCproject.html.

Kulkarni, C., Wie, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., Klemmer, S. R. (2013). Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction, 9*(4), Article 39. http://dx.doi.org/10.1145/2505057

Luo, H., Robinson, A. C., & Park, J. Y. (2014). Peer grading in a MOOC: Reliability, validity, and perceived effects. Available from: *Online Learning Journal, 18*, 1-14. Retrieved from https://olj.onlinelearningconsortium.org/index.php/olj/article/view/429.

Meier, S. L., Rich, B. S., & Candy, J. (2006). Teachers' use of rubrics to score non-traditional tasks: Factors related to discrepancies in scoring. *Assessment in Education: Principles, Policy and Practice, 13*, 69-95. https://dx.doi.org/10.1080/09695940600563512

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). Tuned models of peer assessment in MOOCs. Retrieved from: https://web.stanford.edu/~cpiech/bio/papers/tuningPeerGrading.pdf

Reddy, Y. M. & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435-448. http://dx.doi.org/10.1080/02602930902862859

Staubitz, T., Petrick, D., Bauer, M., Renz, J., & Meinel, C. (2016). Improving the Peer Assessment Experience on MOOC platforms. *In Proceedings of 3rd Annual Learning @ Scale Conference (L@S2016)*. Edinburgh. http://dx.doi.org/10.1145/2876034.2876043

Suen H. K. (2014). Peer assessment for massive open online courses (MOOCs). *The International Review of Research into Open and Distributed Learning, 15*(3). Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/1680/2904.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249-276. http://dx.doi.org/10.2307/1170598

Wang B. (2010). On rater agreement and rater training. *English Language Teaching, 3*(1), 108-112. http://dx.doi.org/10.5539/elt.v3n1p108

Yousef, A. M. F., Chatti, M. A., Wosnitza, M., & Schroeder, U. (2015b). A cluster analysis of MOOC stakeholder perspectives. *RUSC. Universities and Knowledge Society Journal, 12*(1), 74-90. http://dx.doi.org/10.7238/rusc.v12i1.2253 (Original text in Spanish) ■