Peer assessment in MOOCs: The relationship between peer reviewers' ability and authors' essay performance

4

An adapted version of this chapter has been published as:
Huisman, B., Admiraal, W., Pilli, O., van de Ven, M., & Saab, N. (2018). Peer assessment in MOOCs: The relationship between peer reviewers' ability and authors' essay performance. *British Journal of Educational Technology, 49*(1), 101-110. doi:10.1111/bjet.12520

Abstract

In a relatively short period of time, massive open online courses (MOOCs) have become a considerable topic of research and debate, and the number of available MOOCs is rapidly growing. Along with issues of formal recognition and accreditation, this growth in the number of MOOCs being developed increases the relevance of assessment quality. Within the context of a typical xMOOC, the current study focuses on peer assessment of essay assignments. In the literature, two contradicting theoretical arguments can be found: that learners should be matched with same-ability peers (homogeneously) versus that students should be matched with different-ability peers (heterogeneously). Considering these arguments, the relationship between peer reviewers' ability and authors' essay performance is explored. Results indicate that peer reviewers' ability is positively related to authors' essay performance. Moreover, this relationship is only established for intermediate and high ability authors; essay performance of lower ability authors appeared not to be related to the ability of their reviewing peers. Results are discussed in relation to the matching of learners, and instructional design of peer assessment in MOOCs.

Keywords: Peer assessment; Massive Open Online Course; essay performance; ability match

Introduction

Despite their relatively recent introduction, massive open online courses (MOOCs) have become a topic of research in the field of higher education (Raffaghelli, Cucchiara, & Persico, 2015), as well as a topic of scientific and public debate (Kovanović, Joksimović, Gašević, Siemens, & Hatala, 2015). Since the launch of the "Connectivism and Connective Knowledge" MOOC (Downes, 2008), MOOCs became a trend reaching thousands of participants at a time (Evans, Baker, & Dee, 2016). Such large numbers are perhaps not surprising, considering the unrestricted access to university courses for a global audience. The most influential categorization of MOOC pedagogies distinguishes between more connectivist cMOOCs, on the one hand, and more institutionally oriented xMOOCs, on the other hand (e.g., Admiraal, Huisman, & Pilli, 2015; Terras & Ramsay, 2015). Generally speaking, autonomy, interaction, and a construction-oriented teaching approach are central in cMOOCs (Kop, 2011; Toven-Lindsey, Rhoads, & Berdan Lozano, 2015). In contrast, the more institutionally oriented xMOOCs are often characterized by step-by-step learning paths and an emphasis knowledge transfer (Ebben & Murphy, 2014; Rhoads, Sayil Camacho, Toven-Lindsey, & Berdan Lozano, 2015).

As a new form of distance education, MOOCs are in many ways different from traditional university courses. From a research perspective, DeBoer, Ho, Stump, and Breslow (2014) argue that educational variables need to be reconceptualized altogether. For participants, there is usually limited supervision from or direct contact with the teaching staff. Also, assessment procedures are characterized by automated assessment and peer assessment instead of assessment by the teaching staff. Self- and peer assessment - which have been historically used for logistical, pedagogical, metacognitive, and affective benefits - offer promising solutions that can scale the grading of complex assignments in courses with thousands of participants. How to design self- and peer assessment is a challenge in itself as MOOCs have massive, diverse participant enrollment. Within the context of a typical xMOOC, this study focuses on peer assessment of such relatively complex, open-ended assignments, i.e. essay assignments.

Assessment in MOOCs

With the number of available MOOCs rapidly rising, issues of formal recognition and accreditation become increasingly relevant (Lawton & Lunt, 2013). Indeed, several platforms, such as Coursera and EdX, have started to integrate forms of digital 'badges'. This raises important issues such as the reliability of participant identification and the quality of assessment. Regarding the former, several verification methods are being used in a complementary fashion, such as verification via webcams and individual typing-pattern recognition. Such verification methods will undoubtedly continue to develop in the near future. With respect to assessment quality, reliable and valid assessment of participants' learning is required. A practical limitation of having these large numbers of enrolled participants is that alternatives to assessment by teaching staff need to be considered. Not surprisingly, often-occurring forms of assessments in MOOCs are automatic assessment of quizzes and short answer questions, next to self- and peer assessment of more complex, open-ended assignments such as essays. The value of including assessments of participant-generated, open-ended products seems self-explanatory. However, the question which scalable assessment form or process is optimal for such open-ended assignments is not. Different approaches are possible, such as automated essay scoring (AES; e.g., Chauhan, 2014), which come in both supervised and unsupervised variations (Reich, Tingley, Leder-Luis, Roberts, & Stewart, 2015), and human based assessment such as self- and peer assessment. Arguments for the use of peer assessment are twofold. First, peer assessment can be a valid and reliable way to assess student performance (e.g., Cho, Schunn, & Wilson, 2006; Falchikov & Goldfinch, 2000). Second, peer assessment may not only benefit the receiving individual, but may also be beneficial for the peer reviewer him- or herself (Lundstrom & Baker, 2009), since it exposes the peer reviewer to other examples and requires him- or her to actively consider the goals and criteria of the assignment (Flower, Hayes, Carey, Schriver, & Stratman, 1986). In short, both receiving and providing peer assessment can be expected to enhance learning and performance.

Peer Assessment of Essay Assignments in MOOCs

With essay assignments in MOOCs, participants can receive formative feedback from, as well as summative assessment (grading) by multiple peers. The weighted sum of these peer grades usually determines final essay grades, in which self-assessments are occasionally weighted as well. Compared to self-assessments, though, peer assessments might provide a more valid measure of performance. In a recent analysis of three MOOCs, Admiraal, Huisman, and van de Ven (2014) found that self-assessments were biased, and did not explain variance in final exam scores. In contrast, weekly quizzes and peer assessments significantly explained differences in participants' final exam scores. Moreover, research by Cho and colleagues (Cho & MacArthur, 2010; Cho & Schunn, 2007; Cho, Schunn, & Charney, 2006) indicates that assessment by multiple peers can compete with assessment by an expert in terms of reliability (summative), feedback quality (formative), and subsequent improvement by the receiver. Also, in order to get reliable and valid peer feedback and assessments, clear criteria and standards are essential for both authors and reviewers (e.g., Topping, 1998; van Gennip, Segers, & Tillema, 2009), as well as are clear instructions for the provision of feedback (e.g., Gielen & de Wever, 2015). This is an important reason for the inclusion of rubrics in the peer assessment procedure; they explicate the criteria and standards on which the assignment is to be assessed, aiming to simultaneously increase participants' awareness of these criteria and the quality of the provided peer feedback and assessment.

In addition to assessment by multiple peers, and clear standards and criteria, peer assessment might be improved by taking into account the ability of an author and his or her reviewing peers. However, there does not appear to be consensus on how to optimally match authors and reviewers in terms of ability. One the one hand, some authors (e.g., Topping, 2009) argue that learners should be matched with peers of similar ability (homogeneous matching). On the other hand, research by for example Patchan, Hawk, Stevens, and Schunn (2013) suggests that lower ability learners benefit more from assessment by higher ability peers (heterogeneous matching). However, these types of studies on ability matching are generally based on on-campus courses, or at least on

courses in which participants can be expected to be relatively similar in terms of educational background. In open online learning environments such as MOOCS, participants may differ substantially with respect to educational background, ability and motivations. Therefore, a first possible step towards a better understanding of ability matching in open online education is an exploration of how reviewer ability is related to authors' performance, and how author ability and reviewer ability interact in explaining learners' performance. The current study focuses on these questions in the context of a MOOC by Leiden University, launched in 2013.

Research Questions

The central aim of this study is to explore the extent to which peer reviewers' ability is related to authors' essay performance, and to what extent authors' and reviewers' ability interact. Two research questions are formulated. Research question 1 is: "to what extent is peer reviewers' ability related to authors' essay performance?" Research question 2 is: "to what extent does the ability of authors and peer reviewers interact in explaining authors' essay performance?"

Method

The MOOC central to this study is Terrorism and Counterterrorism: Comparing Theory and Practice, organized by Leiden University. It concerns the first run of this particular MOOC, offered in the fall of 2013 via the Coursera platform. The MOOC covered 5 weeks, with an intended workload of 5-8 hours per week.

Participants and Procedure

In total, 26889 participants enrolled for this MOOC. Assessment consisted of five weekly quizzes, two peer reviewed assignments with accompanying self-assessments, and a final exam in the form of a quiz. All five consecutive weeks contained a quiz, and the peer reviewed assignments were scheduled in weeks two and four. The final exam took place in week five (see Table 1 for an overview).

Determination of final grades depended on the track participants choose to follow. In the 'Basic' track, the five quiz scores accumulated to 50%, with the final exam counting as the remaining 50%. In this study, we focus on participants in the 'Advanced' track, which includes the peer reviewed assignments. Here, the quiz scores accumulated to 30%, the two peer reviewed assignments counted for 15% each, and the final exam for the remaining 40%. Within this 'Advanced' track, participants were instructed to review the essays of at least four peers and to perform a self-assessment. A failure to review at least four essays produced by peers and/or submit the self-assessment resulted in a penalty of -20% on the average peer assessment score. This administrative correction is not taken into account in our analysis for two reasons: First, because earlier research showed such self-assessments tend to be biased (Admiraal et al., 2014), and second, because participants' assessment scores then optimally reflect the quality of their submitted work. Self-assessments were done in 94.7% and 95.8% of the cases for assignments 1 and 2 respectively.

Table 1. Chronological overview of assessments (total enrollment = 26889)

Week	Assessment	$N_{(included)}$*	$N_{(total\ submissions)}$
1	Quiz 1	565	5399
2	Quiz 2	564	4077
2	Peer assignment 1	565	842
3	Quiz 3	561	3593
4	Quiz 4	553	3230
4	Peer assignment 2	565	593
5	Quiz 5	544	3014
5	Final exam	540	2988

* = Participants were included when both peer reviewed assignments, *and* at least one of the quizzes was made.

Variables

Quizzes and final exam. The five weekly quizzes were automatically graded, and final scores were based on the best of three possible attempts. Quizzes generally consisted of ten to fifteen multiple choice (MC) questions. For example, one

MC question read "What phrase best explains why terrorism is a contested concept?", with answer alternatives varying from "The enemy of my enemy is my friend" to "One man's terrorist is another man's freedom fighter". Quizzes 1 and 2 slightly deviated from the standard MC question format, both consisting of 9 MC questions plus one open-ended question. These open-ended questions required short answers such as the name of an author, allowing automatic assessment. The final exam consisted of 25 (varying types of) automatically assessed MC questions.

Essay assignments. The two peer reviewed assignments were essay assignments of 600-800 words, excluding references. Each participant was instructed to review at least four peers. A rubric was provided, which allowed for open-ended, freely constructed feedback in addition to every predefined criterion. The predefined criteria of the rubrics slightly differed across the two assignments. Assignment 1 focused on designated terrorist organizations, for which participants were instructed to choose a (in their view) terrorist organization currently not listed as such. The weighted rubric for this assignment included argumentation on chosen examples (40%), argumentation on context (20%), use of sources (30%), and presentation of the essay (10%). Assignment 2 concerned the theoretical assumptions underlying debates on terrorism or counterterrorism, for which participants could choose one of four assumptions to test. The weighted rubric for this assignment included "origin of the claim" (10%), importance of the claim (20%), use of sources (30%), conclusion (30%), and presentation of the essay (10%). Based on these criteria, participants' essay performance was defined as the average score provided by the group of peer reviewers. Within the current study, essay scores were rescaled for interpretation purposes to range between 1 (lowest possible score) and 10 (highest possible score).

Inclusion and Participant Grouping

Participant ability is defined as their average performance on the quizzes made. As such, participants are included in the analysis when they completed both peer reviewed assignments and at least one of the five quizzes. Based on these inclusion criteria, 565 participants are included in this study. In their role as

author, participants are grouped post hoc based on ability, defined as average quiz performance (Avg Q1-Q5). Because of the skewed distribution of scores, a visual binning procedure is used to identify three different ability groups: high ($M = 9.94$, $SD = 0.07$, $N = 237$), intermediate ($M = 9.31$, $SD = 0.34$, $N = 257$), and low ($M = 7.67$, $SD = 0.88$, $N = 71$).

Analyses

To answer the two research questions, hierarchical linear regressions are performed with authors' performance on the second peer reviewed essay assignments (PA2) as dependent variable. For the research question 1, authors' performance on the first peer reviewed essay assignment (PA1) is included as an independent variable in step 1 to control for prior essay performance. Average peer reviewer ability (Avg Q1-Q5) is included as an independent variable in step 2. For research question 2, a similar hierarchical regression analysis is performed while differentiating for the three subgroups of author ability (high, intermediate, and low).

Results

In Table 2, the average quiz score and essay scores are presented, both for the total group of authors and for the three ability subgroups. Scores on peer reviewed essay assignments 1 and 2 are significantly correlated, $r(565) = 0.429$, $p < .001$. However, the mean score for essay assignment 2 ($M = 8.24$, $SE = 2.06$) is lower than the one for essay assignment 1 ($M = 8.75$, $SE = 1.60$), $t(564) = 6.13$, $p < .001$. Apparently, the second essay assignment was more difficult than the first. Further, average quiz performance (ability measure; Avg Q1-Q5) correlates significantly with performance on their first essay assignment: $r = 0.301$, $p < .001$. Thus, authors' ability is moderately correlated to their initial essay performance, before the peer assessment phase.

The central aim of this study is to explore the extent to which peer reviewers' ability is related to authors' essay performance, and to what extent authors' and

reviewers' ability interact. Two research questions were formulated, which are addressed one by one below.

Table 2. Assessment descriptives for author ability subgroups

Assessment	Author ability group							
	Lowest (1)		*Intermediate (2)*		*Highest (3)*		***Total***	
	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N
Avg Q1-Q5	7.67 (0.88)	71	9.31 (0.34)	257	9.94 (0.07)	237	9.37 (0.81)	565
PA1	7.82 (1.87)	71	8.64 (1.66)	257	9.15 (1.28)	237	8.75 (1.60)	565
PA2	7.58 (2.13)	71	8.01 (2.26)	257	8.68 (1.69)	237	8.24 (2.06)	565
Final exam	6.26 (1.93)	67	7.49 (1.73)	244	8.55 (1.13)	232	7.79 (1.71)	543

Peer Reviewers' Ability and Authors' Essay Performance

The ability of peer reviewers appears to be positively related to authors' essay performance ($\beta = 0.13$, $t(563) = 3.37$, $p < .001$, $R^2 = .016$), see Table 3. Thus, while correcting for prior essay performance, the ability of peer reviewers is positively related to authors' performance on a subsequent essay assignment. This effect appears to be small, however (Cohen, 1988).

Interaction Between Authors' and Peer Reviewers' Ability

To assess whether the positive influence of peer reviewers' ability on essay performance varies for authors of different ability levels, regression analyses were performed with the three subgroups of authors: relatively low, intermediate, and high ability, as indicated by their average quiz scores. Indeed, there appears to be an interaction between authors' ability and peer reviewers' ability. Specifically, peer reviewers' ability is positively related to the essay performance of the intermediate ability authors ($\beta = 0.11$, $t(255) = 2.03$, $p = .044$, $R^2 = .013$) and high ability authors ($\beta = 0.22$, $t(235) = 3.56$, $p < .001$, $R^2 = .046$), see Table 4. Here too, however, these effects appear to be small (Cohen, 1988).

Table 3. Regression coefficients for essay performance

Author ability group	Step	Variables included	*B*	*SE*	*β*	*t*	*sig.*
Total	1	Constant	3.40	0.44			
		PA1 score	0.55	0.05	0.43	11.27	.000
	2	Constant	-1.68	1.57			
		PA1 score	0.55	0.05	0.43	11.39	.000
		Peer reviewers' ability	0.55	0.16	0.13	3.37	.001
Low	1	Constant	3.94	1.00			
		PA1 score	0.47	0.13	0.41	3.74	.000
	2	Constant	4.81	3.63			
		PA1 score	0.47	0.13	0.41	3.72	.000
		Peer reviewers' ability	-0.09	0.38	-0.03	-0.25	.802
Intermediate	1	Constant	2.80	0.67			
		PA1 score	0.60	0.08	0.44	7.88	.000
	2	Constant	2.40	2.65			
		PA1 score	0.59	0.08	0.43	7.76	.000
		Peer reviewers' ability	0.57	0.28	0.11	2.03	.044
High	1	Constant	4.80	0.75			
		PA1 score	0.42	0.08	0.32	5.21	.000
	2	Constant	-2.92	2.29			
		PA1 score	0.46	0.08	0.35	5.74	.000
		Peer reviewers' ability	0.79	0.22	0.22	3.56	.000

Note:
$R^2_{Total} = .184$ for step 1, $\Delta R^2 = .200$ for step 2 ($p = .001$) dependent variable: PA2 score
$R^2_{Low} = .169$ for step 1, $\Delta R^2 = .001$ for step 2 ($p = .802$)
$R^2_{Intermediate} = .196$ for step 1, $\Delta R^2 = .013$ for step 2 ($p = .044$)
$R^2_{High} = .104$ for step 1, $\Delta R^2 = .046$ for step 2 ($p < .001$)

Discussion

In this study, we explored how the average ability of peer reviewers relates to authors' essay performance, and to what extent authors' and peer reviewers' ability interact in explaining differences in essay performance. In general, the ability of the reviewing peers was significantly related to authors' essay performance: the higher the ability of peer reviewers, the more authors' essay performance increased. However, this is not the case for all authors: only the essay performance of the (relatively) intermediate and high ability authors is related to peer reviewers' ability, whereas that of lower ability authors is not. Except for this group of relatively low ability participants, this finding supports the idea of matching of MOOC participants with high ability reviewers during peer reviewed essay assignments.

Different explanations are conceivable, which do not necessarily exclude each other. For example, the very ability to utilize received feedback could have an effect on authors' essay performance. This may imply that participants, perhaps especially those of low ability, may benefit from training or guidance on utilizing feedback. Alternatively, and possibly complementary to the former, these findings may indicate that the quality of the provided feedback could be improved. One possible approach here could be to enhance feedback quality by increasing the awareness of different task aspects such as content, structure, style, and to stimulate the provision of more concrete suggestions for revision (e.g., Nelson & Schunn, 2009; van den Berg, Admiraal, & Pilot, 2006a, 2006b). Another approach could be to provide more structured guidance during the feedback process of peer assessment, for example through detailed feedback templates (Gielen & de Wever, 2015).

Implications and Limitations

Certain limitations regarding the current study need to be addressed, and some cautions are in place when interpreting the results of this study. First, the exact mechanism through which peer reviewers' ability is related to the essay performance of intermediate and high ability authors' remains unclear. It is

possible that peer reviewers' ability is related to the quantity or quality of the feedback, and that higher ability authors are better at utilizing this feedback from high ability peers. Since this study does not assess the quantity or quality of peer feedback comments, or the degree to which revisions are done based on received peer feedback (e.g., Patchan & Schunn, 2015), it remains an open question what the exact role of peer feedback has been. Second, this study focuses on received peer assessments. It is possible that the very act of providing peer assessments contributes to participants' learning too (cf. Lundstrom & Baker, 2009), and that providing peer assessment is particularly beneficial for higher ability participants because they tend to more actively consider the assignment goals and criteria (e.g., Flower et al., 1986; Patchan et al., 2013). For future research in online and on-campus education, research on the relation between author and reviewer ability, feedback quality and essay performance seems a fruitful endeavor.

Finally, this study aimed to provide a first exploratory step towards a better understanding of ability matching in open online education. With such first steps however, the degree to which results can be generalized is limited. For one, the available information on the MOOC participants in this study is limited; we have no information with respect to participants' national or educational background, age, or professional occupation. In addition, and potentially related to these variables, it remains unknown whether participants' preference for particular topics, learning activities (i.e. peer assessment), and assignment types (i.e. argumentative texts) influences how they perform peer assessments. In the current study, participants were grouped randomly and not based on such variables. As such, they could be presumed to be relatively evenly distributed over the different ability groups, making them unlikely to confound with the variables used in the analyses of this study. Either way, with respect to future MOOC design and MOOC research, more information on participants could prove valuable. Especially if MOOC platforms would facilitate (quasi-)experimental interventions within MOOC iterations (e.g., A/B testing) or between cohorts of participants, variables such as participants' national or educational background could be interesting matching criteria. This

information on participants should ideally be available a priori, for example through pre-course surveys, in order to purposefully match participants during the peer assessment phase in a MOOC. Another limitation of this explorative study is that only one MOOC was studied, and that the topic of terrorism may be sensitive to participant characteristics such as national background. Hence, future research on peer matching should include MOOCs with different course designs, on different topics and from different platforms, in order to validate the current findings.

Despite these limitations, this empirical study contributes to our knowledge regarding peer assessment in MOOCs. The study provides a first insight into the relationship between the ability of authors and peer reviewers in peer assessment with essay assignments, and gives directions for future research on online peer assessment practices. We believe these findings to be informative for educational developers involved in the instructional design of MOOCs, and hope to instigate future research on peer matching in both open online and on-campus education.

Statement on Open Data

The anonymized data and syntaxes are accessible via the following link: https://osf.io/fv4mw