The Pennsylvania State University

The Graduate School

College of Education

EFFECT OF CHARACTERISTICS OF PEER RATER ON VALIDITY OF PEER ASSESSMENT IN MASSIVE OPEN ONLINE COURSES (MOOC)

A Thesis in

Educational Psychology

by

Xiuyan Guo

© 2015 Xiuyan Guo

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2015
The thesis of Xiuyan Guo was reviewed and approved* by the following:

Hoi K. Suen  
Distinguished Professor of Education  
Thesis Advisor

Pui-Wa Lei  
Associate Professor of Education

Peggy N. Van Meter  
Associate Professor of Education  
Head of the Department of Educational Psychology

*Signatures are on file in the Graduate School
ABSTRACT

The present study investigated the effect of the following two characteristics of peer rater on validity of peer assessment in MOOCs: knowledge of course content measured by the assessment and seriousness in making responsible judgments. The extent of participation of MOOC requirements and the course performance are two indicators being used to identify high quality peer raters who have good knowledge of course content; verified certificate registration and rating time are two indicators being used to identify high quality peer raters who demonstrate seriousness judgments. The author attempts to determine whether high quality peer raters can provide more accurate scores than their counterparts in MOOCs. Our findings yielded clear evidence that a peer rater with higher extent of participation and better course performance, who spend sufficient time in rating peers’ assignment, or who pay the verified certificate fee tend to provide a more accurate score than those with lower extent of participation, poor course performance, who do not spend the necessary amount of time, or who do not pay the verified certificate fee. However, none of them had statistically significant result, and all the effect sizes were very small, showing little practical significance. The limitations of this study due to use of secondary data analyses are discussed. For future research, the author recommends researchers to begin with measuring and investigating some controllable independent variables in a traditional online peer assessment, and then apply the confirmed effective variable in a MOOC environment to collect more accurate and meaningful data.
# TABLE OF CONTENTS

List of Figures .............................................................................................................. v

List of Tables ............................................................................................................... vi

Acknowledgements ...................................................................................................... vii

Chapter 1 Introduction ................................................................................................. 1

1.1 Background .............................................................................................................. 1
1.2 Statement of the Problem ......................................................................................... 3
1.3 Current Study .......................................................................................................... 4

Chapter 2 Literature Review ......................................................................................... 5

2.1. Peer Assessment in Traditional Classrooms and Online Classrooms ................. 5
2.2 Proposed Indicators of High Quality Peer Rater in MOOC ................................. 8
   2.2.1 Content Knowledge ..................................................................................... 11
   2.2.2 Seriousness in Making Responsible Judgments ........................................... 13
2.3 Research Questions and Study Hypotheses ....................................................... 15

Chapter 3 Research Method ....................................................................................... 16

3.1 Participants .............................................................................................................. 16
3.2 Procedure ............................................................................................................... 17
3.3 Instrumentation and Measures ............................................................................. 17
   3.3.1 The Extent of Participation ....................................................................... 17
   3.3.2 The Course Performance ........................................................................... 19
   3.3.3 Certificate .................................................................................................... 21
   3.3.4 Rating Time .................................................................................................. 21
   3.3.5 Peer Assessment ......................................................................................... 22
   3.3.6 Validity ........................................................................................................ 23
3.4 Data Analysis ......................................................................................................... 23

Chapter 4 Findings ..................................................................................................... 25

4.1 The Extent of Participation and Validity of Peer Assessment ............................ 29
4.2 The Course Performance and Validity of Peer Assessment ............................. 30
4.3 Verified Certificate and Validity of Peer Assessment ....................................... 33
4.4 Rating Time and Validity of Peer Assessment ................................................... 33

Chapter 5 Discussion, Implication, Limitations and Future Research .................... 35

5.1 Discussion of the Overall Findings ..................................................................... 35
5.2 Limitation, Implication and Future Research ..................................................... 38
LIST OF FIGURES

Figure 4-1. Histogram of standardized residual before transformation of dependent variable. ................................................................. 26

Figure 4-2. Diagnostic plot for homoscedasticity assumption before transformation of dependent variable ........................................ 27

Figure 4-3. Histogram of standardized residual after transformation of dependent variable ................................................................. 28

Figure 4-4. Diagnostic plot for homoscedasticity assumption after transformation of dependent variable ........................................ 29
LIST OF TABLES

Table 1-1. Regression coefficients, predicting accuracy of peer assessment ........... 31

Table 2-1. Correlations matrix for the measured variables along with their mean and standard deviation. ................................................................. 32
ACKNOWLEDGEMENTS

This thesis would not have been possible without many people. First and foremost, I would like to thank my academic advisor, Dr. Hoi K. Suen. It is yours because of you that I finally found the right focus for my research interests. I greatly appreciate the efforts you put into cultivating me and this thesis.

Dr. Pui-Wa Lei, thank you for the training, mentoring, and support you’ve provided to me. You have always been so encouraging and supportive since my very first day in the Educational Psychology program.

Special thanks must go, too, to my dearest peers Yao Xiong, Deborah Goins, and Xiaoli Jiang. I am so lucky to have you as schoolmates in the Educational Psychology program at Penn State. Thank you for assisting me in every aspect of my life.

Barton K. Pursel, thank you for helping me to get access to the data source and your expertise with data management.
Chapter 1
Introduction

1.1 Background

Since George Siemens invented the first massive open online course (MOOC) with his colleague at University of Manitoba in 2008, MOOC has been generating worldwide interest and lively discussions in higher education (Mackness, Mak, & Williams, 2010). Within such a short period, numerous instructors worldwide have developed and offered MOOCs online at no financial cost. In 2014, there were approximately 1,000 MOOCs available in the United States and 800 from European universities (Bates, 2014). An array of top universities and venture-capital funds invested millions of dollars into the MOOC marketplace. At this time, Coursera is the largest platform provider offering almost half of all the MOOCs. However, there are many other providers (e.g., Udacity, edX, Canvas Network, Open2Study, FutureLearn, Miriada X, France Universite Numerique, and NovoEd) which have mushroomed from different parts of the world. Also, MOOCs are offered in many languages besides English; including Spanish, French, Chinese, Italian, Portuguese, German, Arabic, and Vietnamese. Hundreds of thousands of people of diverse goals, experiences, languages, cultural backgrounds, and motivations have enrolled in MOOCs. At the end of 2014, there were more than 16.8 million students who have enrolled in MOOCs (Shah, 2014).
In a way, MOOCs are a combination of online learning and open educational resources (Bali, 2014). Moreover, MOOCs serve as “the college education revolution” for millions around the world who lack access to quality education despite strong motivation and ability to study (Friedman, 2012).

Still in their infancy, MOOCs show a lot of promise and long-term potential. Many MOOCs are offering or planning to offer participants certificates, credentials, college credits and even college degrees for participation and completion. After examining the quality of MOOCs, the American Council on Education (ACE), a group that represent most of the country’s university and college presidents, has endorsed five MOOCs for credit, and has advised its 1,800 member colleges to feel comfortable in conferring credit on students who complete and pass these courses (Kolowich, 2013). Antioch University became the first U.S. institution to receive approval from Coursera to offer college credit for specified Coursera MOOCs. In October, 2012, Antioch University Los Angeles (AULA) launched a MOOC pilot program in which students were able to enroll in two Coursera courses developed by the University of Pennsylvania: Modern and Contemporary American Poetry, and Greek and Roman Mythology. The AULA/Coursera program has been specifically designed to ensure that students enrolled in a MOOC have a comprehensive and rigorous learning experience. By adding face-to-face interaction with Antioch University faculty and supplemental resources to the existing materials that are a part of the MOOCs, AULA was firmly convinced that these courses are worthy of credit (Antioch University, 2014).

Furthermore, California Senator Darrell Steinberg begins to demand universities to accept MOOCs for credit (Lewin, 2013). Not only do these rapid developments of
MOOCs give the massive studious participants a ray of hope of getting a college degree with little to no cost, but also demonstrate a possible solution to budgetary issues for government.

1.2 Statement of the Problem

Although few will deny the wide popularity and promising benefits of MOOCs, many doubt their pedagogical effectiveness. With its unique online learning environment, reputable instructors and quality course contents alone do not equal to effective learning in MOOCs. The most distinguishable difference between MOOCs and traditional online courses is the teacher-to-student ratio. For each MOOC, there can be as many as over one hundred thousand enrolled students, so instructor-to-student interaction and feedback are fiscally and physically impossible. Furthermore, the teach-learn-assess cycle is broken. As a result, it is difficult for students and instructors to engage in appropriate and in-time feedback for enhanced learning.

Currently, MOOC providers and researchers propose and utilize a variety of approaches to assess learning such as multiple choice testing, on-line discussion forum, automated essay scoring algorithms, and flipped classroom. After examining such various assessment methods, Suen (2014) concluded that peer assessment provides the most universally applicable method to introduce assessment back into the cycle in MOOCs. However, students generally do not trust and value the results of unsupervised peer assessments. If unsupervised and unmonitored peer assessment can be reliable and valid, the results can reattach the teach-learn-assess cycle within MOOCs.
1.3 Current Study

Considering the huge difference of MOOC students’ background and goals, it is not reasonable to expect every student is a serious learner and responsible peer rater. For the purpose for this study, high quality peer rater are defined as being knowledgeable and serious peer rater. Being Knowledgeable is indicated by high level of participation in course activities; and being serious is indicated by spending sufficient time on the peer rating task and having registered for Verified Certificate. The present study aims to investigate the characteristics of knowledgeableness and seriousness of high quality peer raters in MOOC, to explore corresponding indicators of these characteristics, and to examine the effect of these characteristics on the validity of peer assessment by examining whether high quality peer raters can provide more accurate score than their counterparts in MOOCs.
Chapter 2

Literature Review

Since MOOC is still fairly new, there is limited research, and relevant research on its peer assessment is sparse. Additionally, it is justifiable to examine the studies of validity with current peer assessment practices in traditional classrooms and non-MOOC online courses and explore some characteristics of high quality peer rater by looking at the existing literature.

2.1. Peer Assessment in Traditional Classrooms and Online Classrooms

Consisting of peer feedback and peer grading, peer assessment is defined as “arrangement in which individuals consider the amount, level, value, worth, quality or success of the products or outcomes of learning of peers of similar status” (Topping, 1998, p. 250). Peer feedback refers to giving comments on the work or performance of peers; while peer grading involves applying set criteria for assigning grades to peers’ work. Since the 1990s, peer assessments have been widely adopted in traditional classrooms as an “assessment for learning” strategy as well as a strategy for engaging students to be active learners (Cho, Schunn, & Wilson, 2006; Mangelsdorf, 1992; Cheng and Warren 1999; Sadler 1989). It can also reduce the instructors’ heavy load of providing assessment feedback.
Although there is a high degree of agreement among researchers and educators on the benefits of peer assessment for effective learning, this degree of consensus does not appear in the literature regarding the reliability and validity of peer assessment. There has been an apparent misconception about reliability and validity in the literature. Falchikov and Goldfinch (2000) point out that validity is sometimes misreported in the literature as reliability. Based on Falchikov and Goldfinch, reliability is about the consistency and stability of scores. In peer assessment, two types of reliability are generally estimated: inter-rater reliability (i.e., the consistency between different raters) and intra-rater reliability (i.e., the consistency within an individual rater). Validity is the degree of agreement or similarity between the scores assigned by peers and those assigned by instructors. It provides validity evidence based on relationship with other variables (AERA/APA/NCME, 2014). The assignment scores given by peers and given by instructor assess the same constructs, so the relationship between peers’ scores and instructor score for a given assignment provide convergent validity evidence. Since the main concern of most researchers is the degree of agreement between the scores given by peers and by instructors, many studies actually investigated the validity of peer assessment, rather than the reliability (Bouzidi & Jailllet, 2009; Cho, Schunn, & Wilson, 2006; Chang, Tseng, Chou & Chen, 2011; Stefani, 1994; Falchikov, 1986).

Commonly used measures of validity of peer assessment include comparisons of means between peer assigned scores and instructor assigned scores, Pearson’s product moment correlation, intra-class correlation, and concordance percentages between peer assigned scores and instructor assigned scores. Even though a few studies showed significant discrepancies between the peer and instructor assigned scores (Chen, 2010;
Mowl and Pain, 1995; Korman & Stubblefield, 1971; Oldfield & Macalpine, 1995), a major-ity of studies found moderate to high consistency between a peer’s score and a teachers’ score (e.g., Chen & Tsai, 2009; Papinczak, Young, Groves, & Haynes, 2007; Salder & Good, 2006; Tseng & Tsai, 2007; Sung, Chang, Chiou, & Hou, 2005; Tsai & Liang, 2009; Gadbury-Amyot et al, 2003; Billington, 1997; Kelmar, 1993; Stefani, 1994).

Falchikov and Goldfinch (2000) conducted a meta-analysis on the validity of peer assessment from 1959 to 1999 by examining 48 quantitative studies that compare the numerical marks or grades awarded by peers and faculty in traditional classrooms. With a correlation coefficient ranging from 0.14 to 0.99, the mean overall value was $r=0.69$, which indicates strong evidence of average agreement between peer and teacher marks. Subsequently, Li et al. (2015) conducted a meta-analysis of 69 studies from 1999 to 2013 about the validity of peer assessment in online classroom environments. Results indicated a correlation coefficient between peer and teacher ratings that ranged from -0.19 to 0.98 with a moderately strong estimated average Pearson correlation at 0.63. In sum, empirical evidence of moderately high correlations between peers’ scores and instructors’ scores show that both in-person peer assessment and online peer assessment are reasonably valid assessment approaches.

Regarding the effect of peer raters’ characteristics on the validity of peer assessment, previous studies have provided evidence that peer raters with good knowledge of the subject area tend to provide more accurate scores on their peer’s work (Bouzidi & Jailet, 2009). In addition to knowledge in the subject matter, peer raters need to put forth serious efforts when participating in peer assessment (Li et al., 2015; Hanrahan & Isaacs, 2001). Therefore, sufficient content knowledge and seriousness of
providing fair and accurate rating are two important characteristics of high quality peer raters in traditional peer assessment. For content knowledge, the commonly used measurement is objective test scores of the same subject matter rated by peer assessment such as final examination scores or unit quiz scores (Bouzidi & Jaillet, 2009). However, very few, if any, studies have explored observable indicators of students’ seriousness in making responsible judgment on their peer’s work. So far, no study has examined the effect of the two characteristics on validity of peer assessment in MOOC’s unique environment.

2.2 Proposed Indicators of High Quality Peer Rater in MOOC

Because MOOCs are still budding, there are very few empirical studies of validity of peer assessment in MOOC. Tsai (2013) examined the validity of peer assessment in a MOOC named “History of the World since 1300” offered by Princeton University on the Coursera platform in 2012. In this study, 170 peer graded assignments were randomly selected, and the average of a group of teaching assistants’ scores was use as the ‘gold standard’ of accuracy. A significantly positive correlation between peer’s scores and corresponding teaching assistants’ scores was found with $r = 0.42$. Another study reported a stronger positive correlation between instructor scores and the median of peer scores in a Coursera MOOC named Maps and the Geospatial Revolution offered by the Pennsylvania State University in 2013, with $r = 0.619$ (Luo, Robinson & Park, 2013). It is the same MOOC course used in current study, but offered by different time. The
current study used the data gathered in the 2014 offering, which is one year later than Lou, Robinson and Park’s study.

In MOOCs, peer assessment functions not only as a formative pedagogical tool, but it also can be a main source of summative assessment. Thus, it is essential that peer assessment has high validity in MOOCs for a meaningful learning experience. The most distinguishable difference between peer assessment in MOOCs and peer assessment in traditional classrooms is peer raters’ background and goals. In traditional classrooms, most of the students take a particular course not only for obtaining content knowledge, but also for getting credits, thus most of them will put a lot of effort into that course. Plus, most of the traditional courses have prerequisites (e.g., a lower level course in related area), which have to be satisfied for registration. Thus, students in the traditional classrooms have similar prior content knowledge of a given course and have similar goals.

However, compared to students in traditional face-to-face classroom and traditional online classroom, there are many types of participants with diverse backgrounds and goals in MOOCs. There is no prerequisite for MOOCs registration, everyone can take it. Among the participants in a MOOC, some may have already gotten Ph.D. degrees in related area, while others may have no formal education at all; some participants enroll in a MOOC purely out of curiosity about the innovative teaching format, while others want to explore a specific interest in the chosen topic of a course. There are also MOOC participants who are enrolled to pursue further education for credit or certificates at reduced costs. As a result, the amount of effort they put in learning
content knowledge in a given MOOC as well as their seriousness in making responsible judgment of their peers’ work vary dramatically.

Because peer assessment is a process in which peer raters apply criteria to assign grades to the work of peers based on peer rater’s understanding of the content, it has two requirement from peer rater: (1) A rater should have good knowledge of the course content evaluated by the assessment and (2) should demonstrate seriousness in making responsible judgments. In order to give accurate judgment of peers’ work, peer raters need to have good content knowledge related to the work under assessment, otherwise they cannot tell what is correct and what is wrong, let alone specifying the origin of mistakes and providing valuable feedback. In other words, good content knowledge is the most important and basic requirement for making a valid peer assessment. Without it, all the judgments are no more than guessing, and meaningful feedback is impossible.

Another important influential characteristic of peer raters for validity of peer assessment is seriousness in making responsible judgment. If a peer rater with good content knowledge does not treat peer assessment seriously, he or she may just randomly assigns a score to peers’ work without reading it carefully, or casually writes down some meaningless feedback. Therefore, peer raters’ seriousness in making responsible judgments is also necessary for accurate peer assessment.

Although implementation of strategies such as non-anonymous versus anonymous raters, paper-based over computer assisted methods of rating, and student-constructed rating rubrics vs. instructor-constructed rubrics have been shown to be successful in improving the validity of peer assessment in traditional classroom settings, these same strategies are not applicable in the MOOC environment (Li et al., 2015). Peer assessment
is conducted by peer raters, so its validity heavily depends on peer raters’ quality. Due to the unique environment in MOOCs, peer raters’ quality varies dramatically. In order to obtain high validity of peer assessment in MOOCs, the crux of the matter is to have high quality peer raters. As mentioned earlier, for the purpose of this study, a peer rater is regarded as having high quality if the individual (1) has a good knowledge of the course concerned by the assessment and (2) demonstrates a seriousness in making responsible judgments. Understanding the effect of the characteristics of high quality peer rater can help to develop valid peer assessment structures in providing meaningful learning experiences through MOOCs.

2.2.1 Content Knowledge

Generally, MOOCs have high quality content and instructional design because prominent professors from top universities are often recruited to participate in the online platform. In many cases, MOOCs simulate a full course at elite universities with a minimum of two to three hours of lecture per week, supplemental readings, frequent homework assignments, and several quizzes. Also, MOOCs are provided in learner-friendly teaching formats. In traditional classrooms, lectures tend to be lengthy and given in a single time frame. However, MOOC lecture videos are divided into many short, modular units of eight to twelve minutes; each video represents a coherent concept. Students have the options to watch the lecture segments as many times as they want. Furthermore, due to technologically sophisticated innovations, students can even
slow down the lecture to 0.5 times the actual speed, listen in actual time, or speed them up by a factor of 1.25 or 1.5.

In order to encourage interaction between students and instructor and collaboration among students, MOOCs usually provide a question-and-answer forum. Inside the forum, students can post their questions and comments about the content of the course. The forum also allows other students to respond to the questions. Some instructors draw a sample of questions and comments and provide their own responses; while other instructors use teaching assistants to identify common concerns and then post comments on those shared concerns while making corrections to some common mistakes or misunderstandings. Additionally, MOOCs provide short quizzes in the form of multiple choice questions with answers and explanations popping up after the students have submitted their answers to help students monitor their understanding of the material.

Considering the generally high quality content and knowledgeable instructors in MOOCs, one can expect to gain sufficient knowledge by completing course tasks. It is reasonable to assume that a peer rater who completes a high proportion of course tasks and has good course performance would have gained a higher level of content knowledge than one who completes a low proportion of course tasks and has poor course performance. Therefore, one criterion for high quality peer raters with good knowledge is a high extent of participation of the course tasks which may include watching videos, completing assignments, posting questions or answers in the forum, and completing peer assessments. The other criterion for high quality peer raters with good knowledge is good course performance which may include peer raters’ test scores, and assignment scores.
2.2.2 Seriousness in Making Responsible Judgments

To be a high quality peer rater, mastering relevant knowledge is necessary but insufficient. Although peer assessment is a practical method of providing feedback in a large classroom setting, it is a time consuming and demanding task that requires effort of the raters. Without seriousness of purpose, peer raters in MOOCs may not fully apply their knowledge and skills when assessing their peers’ assignments.

Seriousness, to some extent, can be indirectly indicated by economic investment. To promote seriousness for students, some MOOCs from Coursera begin to offer two kinds of certificate options. Most MOOCs offer a free “Certificate of Completion” for those who complete the course. Some instructors may have some additional completion requirements for receiving the free “Certificate of Completion.” However, such free certificates of completion are regarded as notification of completion for students. For students who need to show valid proof of their MOOC course and achievement to their prospective employer, they can earn the status of being “Verified Certificated” on signature track by completing a course and paying a certification fee. In such a case, a MOOC’s Signature Track securely links a student’s coursework to his or her identity, thus, confirming that the particular paying student is indeed the individual who completed the MOOC work. In addition to the course name and instructor signature, the “Verified Certificate” features the logo of the partner institution offering the course, a statement attesting to a student’s confirmed identity, and a certification URL that allows others to check the certificate’s authenticity.

The Verification Certificate fee also varies from course to course ranging between
Most MOOCs offered by Coursera charge $49 for their verified certificates. In some developing countries, such as India, $49 can be a considerable amount of money relative to their per capita GDP, which is $1,498.9 for 2013 (World Bank Open Data, 2014). The students who pay the certificate fee can be considered more serious for most tasks in MOOC than those who do not. For a comparable analogy, a person may plan to exercise every day either at home with no fee charged or at a gym paying a monthly membership fee. It would be generally believed that that person is more likely to keep on working out in the latter situation since the membership fee may well work as an incentive for him to exercise on a regular basis. Hence, paying a fee may be an indicator that a person is more serious about accomplishing a task.

Paying the fee for verified certificate reflects the seriousness of MOOC students, and it is reasonable to assume that those students are also serious and responsible peer raters. However, a more direct indicator for serious peer raters is whether they spend enough time rating their peers’ assignments. If a peer rater only spend several seconds rating an assignment, we assume that he or she is assigning a number without carefully reading the peer’s work. Consequently, apart from the fee payment for verified certificate, another criterion for serious peer raters can be indicated by the length of time spent on rating peers’ assignments.
2.3 Research Questions and Study Hypotheses

According to the three aforementioned indictors, we divide MOOCs peer raters into different groups (e.g., the verified-certificate group and the non-verified-certificate group, the necessary -rating-time group, and the insufficient-rating-time group) to compare the validity between these groups and examine the relationship between extent of participation and validity of peer grading. The present study attempts to determine whether high quality peer raters tend to provide more accurate scores for peer assessment in MOOCs. The study hypotheses are as follow.

*Hypothesis 1:* The extent of participation of peer raters is positively correlated with the validity of their peer assessment in MOOC.

*Hypothesis 2:* The course performance of peer raters is positively correlated with the validity of their peer assessment in MOOC.

*Hypothesis 3:* The validity of peer assessment in the verified-certificate group is higher than those in the non-verified-certificate group in MOOC.

*Hypothesis 4:* The validity of peer assessment for peer raters who spent necessary time in peer rating is higher than those did not in MOOC.
Chapter 3

Research Method

This is a secondary data analysis. All the data used have been automatically collected by the Coursera system and have been exported and stored at the Pennsylvania State University Wiki Space managed by the Office Teaching and Learning with Technology.

3.1 Participants

MOOC data were collected from Maps and the Geospatial Revolution MOOC offered by the Pennsylvania State University on the Coursera platform in 2014 (Robinson, Kerski, Long, Luo, DiBiase, & Lee, 2015). The total enrollment in that course was 25,839. Out of the 25,839 students enrolled, 1,336 had submitted an assignment named “storytelling with map”, and 1,227 students had completed peer ratings of this assignment.

For the purpose of this study, the instructor of this course randomly selected about 10%, or 131 submitted assignments to rate. These 131 assignments had been rated by 591 peer raters in this MOOC. The study is based on the analysis of these 591 peer raters. The identities of the peer raters were anonymous.
3.2 Procedure

The Maps and the Geospatial Revolution course was a five-week MOOC focusing on geospatial technology, making maps, and analyzing geographic patterns using the latest tools. In the last week of the course, students were expected to submit an assignment about applying the knowledge learned in this course to create a map in a storytelling context. In addition, students were asked to rate six completed assignments submitted by peers for the storytelling project based on an instructor-designed rubric and to provide open-ended written feedback to these peers. For the purpose of this study, the instructor randomly selected and graded 131 students’ assignments which were used as the ‘gold standard’ of accurate assessment results. When grading the random selection of 131 assignments, the instructor had no information about students’ identities, peer awarded scores, or any other performances in the course.

3.3 Instrumentation and Measures

3.3.1 The Extent of Participation

The main tasks of this MOOC included watching videos, taking quizzes, doing assignments, completing peer assessment, and posting questions or offering answers on the course forum. Because every participant in this study had taken the quizzes, submitted assignment, and completed peer assessment, there were two meaningful components of the extent of participation: the number of lecture videos watched and the number of forum posts.
The number of lecture videos watched

Most of the content knowledge of this MOOC was conveyed through totally 10 lecture videos, with two provided for each week. Every time a student clicked and opened a video, it was recorded in the Coursera database. Although the number of times a student opening a lecture video does not necessarily mean that the student watched the video in its entirety, this record gives us a general idea about students’ activity of watching lecture videos. MOOC video records showed that some students opened a single lecture video more than 50 times, and it is unlikely that the student carefully watched thoroughly the same video that many times. We will discuss the limitation of using this video activity record later in the discussion section. Although the weekly record provided by Coursera cannot tell us who watched lecture videos from beginning to end, it can inform us who did not watch all the videos. In other words, if the weekly record for a participant is zero in a given week, we have 100% confidence to say that this individual did not watch any videos. Thus, in this study, we counted this person as “0.” If the weekly record is 1, we counted the student as having watched one video and counted them as “1”. Likewise, if the weekly record is 2, we counted as “2.” However, if the weekly record was greater than 2, we still counted as “2” because there were only two videos posted each week and the higher number of times a student watched the videos (i.e. 50) does not necessarily mean that the student watched the video so many times from start to finish for meaningful learning. Thus, using this coding method, the total number of videos watched by a student over the 5-week period is between 0 and 10.
**The number of forum posts**

The number of forum post was used as another indicator of students’ participation and was assumed to help student gain content knowledge. We used weekly data provided by Coursera about the number of times a student posted a question or an answer in the forum. Since some students also posted irrelevant messages on the forum, the number of forum post measures the level of participant of students but not the quality of participation. Therefore if a student had more than two posts in one week, we still counted it as “2” assuming one post for asking a question and the other post for providing an answer. The counting method is similar to the one we used to count the number of videos watched per week. Thus, the total number of forum posts ranged from 0 to 10.

### 3.3.2 The Course Performance

In this MOOC, the course performance of each peer rater was evaluated from two main aspects: the test score and assignment score.

**Test score**

There were four quizzes and one final examination in this MOOC to help student master the content knowledge. For each quiz, the students had unlimited attempts to complete the quiz. However, students were only allowed one attempt for the final examination. In general, it appears that students with lower test scores tend to attempt the same quiz multiple times. Since the four quizzes are all in the format of multiple
choice questions, it is reasonable to believe that those students who attempted the same quiz multiple times were randomly guessing the answers in order to find the correct answers. Therefore, to account for high quiz performances due to random guessing, we used the quiz score from the first attempt because it was most likely the person’s true score. In this study, we added up each first tried quiz scores and the final examination score for each student and used that total score as one of the indicators of their content knowledge. The scale for the four quizzes and the final examination were 17, 15, 18, 15 and 50 respectively. Thus student test performance scores ranged from 0 to 115.

Assignment score

The Maps and the Geospatial Revolution MOOC had only one assignment. It was about applying the knowledge learned in this course to draw a map within a story context. The assignment was graded on four criteria: 1) Does this map tell a complete story? 2) Is the story that this map tells presented in a compelling way? 3) Is the map designed in a way that reflects the use of best practices in cartographic design and geospatial analysis? 4) Does this map have an aesthetic look and feel that reinforces the objectives of the story it tells? The scoring scale for each criterion was from 1 to 9 points, with 1 indicating “very strongly disagree,” and a 9 indicating “very strongly agree.” The assignment total scores ranged from 0 to 36.

For each of the four aspects mentioned above, the extent of participation of a students was calculated by dividing the obtained score by the maximum score of that aspect. For example, if the number of lecture videos watched by a student is 9, the extent
of participation of watching lecture videos for a student was calculated as dividing 9 by 10 (i.e. the maximum number of lecture videos being watched in our record) = 0.9. Similarly, we calculated the extent of participation for forum posts, quiz performances, and assignments. The average of the extent of participation from these four aspects was the overall extent of participation of a student.

3.3.3 Certificate

At the beginning of the course, all students were asked about whether they would like to register for a verified certificate named signature track by paying the registration fee of $49. For those who paid verified certificate fee, the ‘certificate’ variable was coded as “1” and those who did not, the variable was coded as “0”. Thus, students were categorized into two groups: the verified-certificate group and the non-certificate group.

3.3.4 Rating Time

For each peer assessment, Coursera provided data about when a student opened an assigned peer’s work and the time in which he or she submitted their peer assessment in Unix time stamps. We used the duration between the two time stamps as an indicator of the amount of time a student spent in rating that peer’s assignment. The challenge was in determining whether the student was fully on-task and engaged in rating the peer’s assignment during this time period. The data showed that the durations of time stamps for some students were more than 90 hours, which most probably indicate that these
student had been off-task and were not fully focused on rating a particular assignment during this period. Although it is difficult to ascertain the exact time a student spent in peer rating, it is feasible to determine whether the students had spent sufficient time in peer rating. Because the instructor spent at least three minutes to grade a single assignment, we regarded the students who spent less than three minutes as off-task and less than serious peer raters by using the instructor’s time spent as the standard cut-off. Therefore, the peer raters were classified into two groups: the necessary-rating-time group, which included all students who had spent at least three minutes on the rating task; and insufficient-rating-time group, which included all students who had spent less than three minutes on the rating task. The former was coded as “1” and the latter as “0.”

3.3.5 Peer Assessment

Each student was also required to complete the peer assessment of six assignments submitted by other students based on the above criteria. The final score of the assignment was the median of peer assigned scores if the student also complete self-assessment. According to the course syllabus, for those who did not complete self-assessment, there was one unit score penalty, that is, the final score of the assignment equals the median of peer assigned scores minus one. In this study, all 591 participants completed their peer assessment; therefore, completion of peer assessment was not a meaningful predictor here.
3.3.6 Validity

For the purpose of this study, the degree of inaccuracy of each peer rater was calibrated in the following manner: The squared distance between the peer rater score and the instructor score for each assignment was calculated. The average of all the squared distance measures for a given peer rater was taken as the degree of inaccuracy of that rater, as shown in the Equation for Distance index below (Equation 1):

\[
\text{Distance index} = \frac{\sum_{i=1}^{n} (S_i - I_i)^2}{n}
\]  

(1)

In the equation, \( n \) is the number of assignments a peer rater graded, \( S_i \) is the score awarded by this peer rater on a particular assignment, \( I_i \) is the score awarded by the instructor on the same assignment. In this study, \( n \) depends on how many assignments assigned to a student to grade were also graded by the instructor, thus, this number varies from one to six since each student was expected to grade six peer assignments. The distance index obtained from this equation tells us, on average, how inaccurate or inconsistent a peer rater is when using instructor’s score as the gold standard. A smaller value indicates less inconsistence between a peer rater’s score and the instructor’s score, and thus indicates higher validity of peer assessment. Distance index of zero means that each of the score a peer rater gives is exactly the same as the instructor’s score.

3.4 Data Analysis

To examine the effect of all the four indicators of high quality peer raters (i.e., extent of participation, course performance, verified certificate registration, and rating
time) on peer assessment validity, a multiple linear regression analysis was conducted with validity index as criterion variable. Because there are two components of the extent of participation (i.e., the number of lecture videos watched and the number of forum posts) and two component of course performance (i.e., total score of four quizzes and final examination and score of the single assignment), we included these four components as predictors, instead of the overall extent of participation and the overall course performance, in this multiple linear model. Thus, there are totally six predictors in the model: verified certificate registration, rating time spending on rating tasks, the number of lecture videos watched, the number of forum posts, total score of four quizzes and final examination, and score of the single assignment.
Chapter 4

Findings

Due to the violation of the normality and homoscedasticity assumptions, as indicated by the diagnostic plots of distributions of standardized residuals (see Figure 4-1 and Figure 4-2), we converted our criterion variable (i.e. distance index values) to percentiles: The highest quality peer rater who had smallest distance index value would have a percentile score of 99; while the lowest quality peer rater who had largest distance index value would have a percentile score of 1. Then the probit of each percentile value was used to meet the normality and homoscedasticity assumption. Figure 4-3 and Figure 4-4 showed the effectiveness of this conversion. The overall model is not statistically significant at alpha of 0.05, $F (6, 584) = 0.822$, $R^2 = 0.008$, $p = 0.553$. 
Figure 4-1. Histogram of standardized residual before transformation of dependent variable
Note. Dependent variable is raw values of distance index; independent variable is verified certificate registration, rating time spending on rating tasks, the number of lecture videos watched, the number of forum posts, total score of four quizzes and final examination, and score of the single assignment.
Figure 4-2. Diagnostic plot for homoscedasticity assumption before transformation of dependent variable.

Note. Dependent variable is raw values of distance index; independent variable is verified certificate registration, rating time spending on rating tasks, the number of lecture videos watched, the number of forum posts, total score of four quizzes and final examination, and score of the single assignment.
Figure 4-3. Histogram of standardized residual after transformation of dependent variable

Note. Dependent variable is raw values of distance index; independent variable is verified certificate registration, rating time spending on rating tasks, the number of lecture videos watched, the number of forum posts, total score of four quizzes and final examination, and score of the single assignment.
Figure 4-4. Diagnostic plot for homoscedasticity assumption after transformation of dependent variable.

Note. Dependent variable is raw values of distance index; independent variable is verified certificate registration, rating time spending on rating tasks, the number of lecture videos watched, the number of forum posts, total score of four quizzes and final examination, and score of the single assignment.

4.1 The Extent of Participation and Validity of Peer Assessment

The first hypothesis is that the extent of participation of peer raters is positively correlated with the validity of their peer assessment in MOOC. Results from this MOOC does not support this hypothesis. For the two components of the extent of participation:
the number of lecture videos watched and the number of forum posts, neither of their regression coefficients is statistically significant at 0.05 level (see Table 4-1).

Although both two components are not statistically significant indicators of validity of peer assessment in this MOOC, within the sample in this study, there are weak positive correlations between validity of peer assessment and the number of lecture videos watched and the number of forum posts. And the number of lecture videos watched has a statistically significant positive zero-order correlations with the converted distance index, \( r=0.077, \ p=0.031 \). More details are provided in Table 4-2. However, the correlation matrix shows that two component variables are significantly positively correlated with each other, \( r=0.310, p < 0.001 \). This result is consistent with our expectation as they are components of the extent of participation.

### 4.2 The Course Performance and Validity of Peer Assessment

The second hypothesis is that the course performance of peer raters is positively correlated with the validity of their peer assessment in MOOC. Results from this MOOC does not support this hypothesis. For the two components of the course performance: total score of four quizzes and final examination and score of the single assignment, neither of their regression coefficients is statistically significant at 0.05 level (see Table 4-1). Although both of the two components are not statistically significant indicators of validity of peer assessment in this MOOC, within the sample in this study, there are weak positive correlations between validity of peer assessment and total score of four quizzes and final examination as well as score of the single assignment. However, the correlation
matrix shows that the two component variables are significantly positively correlated with each other, \( r=0.274, p < 0.001 \). This result is consistent with our expectation as they are components of the extent of participation.

Table 4-1. Regression coefficients, predicting accuracy of peer assessment

<table>
<thead>
<tr>
<th>Model</th>
<th>B</th>
<th>Std.err</th>
<th>( \hat{\beta} )</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Constant</td>
<td>-.216</td>
<td>.124</td>
<td>-1.748</td>
<td>.081</td>
</tr>
<tr>
<td></td>
<td>NO.L</td>
<td>.031</td>
<td>.017</td>
<td>.080</td>
<td>1.754</td>
</tr>
<tr>
<td></td>
<td>NO.F</td>
<td>-.003</td>
<td>.015</td>
<td>-.009</td>
<td>.553</td>
</tr>
<tr>
<td>2</td>
<td>Constant</td>
<td>-.536</td>
<td>.346</td>
<td>-1.551</td>
<td>.121</td>
</tr>
<tr>
<td></td>
<td>NO.L</td>
<td>.030</td>
<td>.017</td>
<td>.076</td>
<td>1.754</td>
</tr>
<tr>
<td></td>
<td>NO.F</td>
<td>-.007</td>
<td>.015</td>
<td>-.022</td>
<td>-.486</td>
</tr>
<tr>
<td></td>
<td>TS</td>
<td>.002</td>
<td>.003</td>
<td>.024</td>
<td>.553</td>
</tr>
<tr>
<td></td>
<td>AS</td>
<td>.006</td>
<td>.008</td>
<td>.032</td>
<td>.742</td>
</tr>
<tr>
<td>3</td>
<td>Constant</td>
<td>-.652</td>
<td>.437</td>
<td>-1.494</td>
<td>.136</td>
</tr>
<tr>
<td></td>
<td>NO.L</td>
<td>.029</td>
<td>.017</td>
<td>.074</td>
<td>1.696</td>
</tr>
<tr>
<td></td>
<td>NO.F</td>
<td>-.008</td>
<td>.015</td>
<td>-.023</td>
<td>-.510</td>
</tr>
<tr>
<td></td>
<td>TS</td>
<td>.002</td>
<td>.003</td>
<td>.025</td>
<td>.557</td>
</tr>
<tr>
<td></td>
<td>AS</td>
<td>.006</td>
<td>.008</td>
<td>.031</td>
<td>.715</td>
</tr>
<tr>
<td></td>
<td>RT</td>
<td>.130</td>
<td>.299</td>
<td>.018</td>
<td>.435</td>
</tr>
<tr>
<td>4</td>
<td>Constant</td>
<td>-.649</td>
<td>.437</td>
<td>-1.486</td>
<td>.138</td>
</tr>
<tr>
<td></td>
<td>NO.L</td>
<td>.028</td>
<td>.017</td>
<td>.073</td>
<td>1.650</td>
</tr>
<tr>
<td></td>
<td>NO.F</td>
<td>-.008</td>
<td>.015</td>
<td>-.023</td>
<td>-.515</td>
</tr>
<tr>
<td></td>
<td>TS</td>
<td>.002</td>
<td>.003</td>
<td>.023</td>
<td>.532</td>
</tr>
<tr>
<td></td>
<td>AS</td>
<td>.006</td>
<td>.008</td>
<td>.031</td>
<td>.714</td>
</tr>
<tr>
<td></td>
<td>RT</td>
<td>.133</td>
<td>.299</td>
<td>.019</td>
<td>.444</td>
</tr>
<tr>
<td></td>
<td>VCR</td>
<td>.040</td>
<td>.109</td>
<td>.015</td>
<td>.364</td>
</tr>
</tbody>
</table>

Note. NO.L= the number of lecture videos watched; NO.F= the number of forum posts; TS=test score; AS=assignment score; RT= rating time (0 = the insufficient-rating-time group; 1= the necessary-rating-time group); VCR= verification certificate registration (0= 
the non-verified certificate group; 1 = the verified-certificate group); $B =$ unstandardized regression coefficients; $\beta =$ standardized regression coefficients.

$N = 591$

Table 4-2. Correlations matrix for the measured variables along with their mean and standard

<table>
<thead>
<tr>
<th></th>
<th>NO.L</th>
<th>TS</th>
<th>AS</th>
<th>NO.F</th>
<th>RT</th>
<th>VCR</th>
<th>CDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO.L</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TS</td>
<td>.150***</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AS</td>
<td>.119**</td>
<td>.274***</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO.F</td>
<td>.310***</td>
<td>.273***</td>
<td>.225***</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT</td>
<td>.137***</td>
<td>.040</td>
<td>.086*</td>
<td>.107**</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VCR</td>
<td>.116**</td>
<td>.087*</td>
<td>.031</td>
<td>.065</td>
<td>-.007</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>CDI</td>
<td>.077*</td>
<td>.039</td>
<td>.043</td>
<td>.016</td>
<td>.029</td>
<td>.025</td>
<td>-</td>
</tr>
<tr>
<td>Mean</td>
<td>7.190</td>
<td>101.217</td>
<td>28.509</td>
<td>3.92</td>
<td></td>
<td></td>
<td>-.005</td>
</tr>
<tr>
<td>SD</td>
<td>2.484</td>
<td>13.469</td>
<td>5.216</td>
<td>2.829</td>
<td></td>
<td></td>
<td>.968</td>
</tr>
</tbody>
</table>

Note. NO.L= the number of lecture videos watched; TS=test score; AS=assignment score; NO.F= the number of forum posts; RT= rating time (0= the insufficient-rating-time group; 1= the necessary-rating-time group); VCR= verification certificate registration (0= the non-verified certificate group; 1= the verified-certificate group), CDI= converted distance index (i.e., probit value).

$N = 591$. 


4.3 Verified Certificate and Validity of Peer Assessment

The second hypothesis is that the validity of peer assessment in the verified-certificate group is higher than those in the non-verified-certificate group in MOOC. Using paying verified certificate fee as the cut-off standard led to 96 peer raters in the verified-certificate group being coded as “1,” and 495 peer raters in the non-verified-certificate group being coded as “0.” The result shows that peer raters in the verified-certificate group tend to provide more accurate scores of peer assessment than those in non-verified-certificate group in this sample, but the result is not significant at 0.05 level, $B=0.040$, $t(584) = .364$. $p = .716$. So, there is not sufficient evidence to support the claim that the validity of peer assessment in the verified-certificate group is higher than those in the non-verified-certificate group in the population, and the second hypothesis is not supported by the data in this MOOC.

4.4 Rating Time and Validity of Peer Assessment

The third hypothesis is that the validity of peer grading for peer raters who spent the necessary amount of time in peer rating is higher than those who did not. As explained earlier, the instructor needed at least three minutes to complete grading one assignment. So we used the average rating time of three minutes as a cut off to classify the students into two groups: the necessary-rating-time group ($n=11$) being coded as “1,”
and the insufficient-rating-time group (n=580) being coded as “0.” The result shows that peer rater in the necessary-rating-time group tend to provide more accurate scores of peer assessment than those in the insufficient-rating-time group holding other independent variables constant, but the difference is not significant at .05 level, $B = .133, t(584) = .444, p = .657$. Therefore, the third hypothesis is not supported by the data of this MOOC, and we cannot conclude that the students who spent the necessary amount of time in grading their peers’ assignment tend to provide a more accurate score than those who did not.
Chapter 5
Discussion, Implication, Limitations and Future Research

5.1 Discussion of the Overall Findings

The present study investigated the effect of two characteristics (i.e., good knowledge of the course content evaluated by the assessment and seriousness in making responsible judgments) of peer rater on the validity of peer assessment in MOOCs. The extent of participation and course performance of peer raters were used as two indicators of content knowledge. The extent of participation comprised two component variables: the number of lecture videos watched, and the number of forum posts, while course performance included total score of four quizzes and final examination as well as score of the single assignment. Verified certificate registration and rating time were two indicators being used to identify high quality peer raters with seriousness in making responsible judgement, and to classify them into different groups. Payment of the $49 verified certificated fee was used as the criterion to classify peer raters into the verified-certificate group vs. the non-verified certificate group. Three minute, which was the time the instructor needed to grade one assignment, was used as the cut-off to classify peer raters into a necessary-rating-time group and an insufficient-rating-time group. The author attempted to determine whether high quality peer raters can provide more accurate score than their counterparts in MOOCs.

The findings yield clear evidence that a peer rater with higher extent of participation and better course performance tend to provide a more accurate score (closer
to an instructor’s score) within the sample analyzed. The zero-order correlations among
four component variable are stronger than the correlations between each aspect and
validity of peer assessment, which indicated that the students who watched more lecture
videos tend to have higher score in quizzes, final examination and assignment, and tend
to participate more on forum discussion, and thus have better content knowledge to
enable them provide a more accurate score on peer assessment. However, none of four
component variables of the extent of participation and course performance is a significant
predictor of for accuracy of peer assessment when controlling other variables as constant.
Among the four variables, only the number of lecture videos watched is significantly
positively correlated with validity index, but the effect size is very small. This can be
partly explained by low reliabilities of measurement. For example, because the number
of lecture videos watched in this study was the number of lecture videos opened, there is
insufficient information to ascertain that students carefully watched the video in its
entirety after opening the video. In an attempt to account for this uncertainty, this study
combined the record of two or more videos watched per week into one category.

Although the results show that peer raters who paid the verified certificate fee
tend to provide more accurate scores than those who did not pay the fee in this sample,
paying the verified certificate fee is not a significant predictor for accuracy of peer
assessment. However, it is interesting to note that the extent of participation of the
verified-certificate group is significantly higher than the non-verified-certificate group.
Since the normality assumption of independent t-test is violated, Mann-Whitney U test
was used. The mean rank of the extent of participation for the verified-certificate group
was 338.24, while the mean rank of the extent of participation for non-verified-certificate
group was 287.81, Mann-Whitney U =19704.50, Z= -2.66, p < 0.01. Similarly, course performance of the verified-certificate group is significantly higher than the non-verified-certificate group. The mean rank of course performance for the verified-certificate group was 331.10, while the mean rank of the extent of participation for non-verified-certificate group was 289.19, Mann-Whitney U =20390.50, Z= -2.20, p < 0.05.

The results demonstrated that paying the fee for verified certificate is an importantly influential factor of the extent of participation and course performance; although not of accuracy level of assessing peers’ work. This indicates that students who paid the verified certificate fee are more serious to complete the course and master the content knowledge, but it does not necessarily mean that the paying raters are more serious in providing more accurate judgment for their peers’ assignment. One possible explanation is that in this Maps and the Geospatial Revolution MOOC, the requirements of gaining a verified certificate only include watching lecture videos, taking quizzes and examination, doing assignment, and finishing peer assessment. As long as the students finished the peer assessment, students met the requirement regardless of the quality or accuracy of their assessment. Since there was no reward for high quality peer assessment nor penalty for the low quality or irresponsible peer assessment, there might not be enough incentive for the student to make an effort to provide quality peer ratings.

Also, with the sample studied, peer raters who spent sufficient time in rating peers’ assignment tend to give a more accurate score on the peers’ work ($M$ = -0.001, $SD$ = 0.975) than those who did not spend necessary time and show little will in making a responsible judgment ($M$ = -0.212, $SD$ = 0.459). But the difference is not significant at 0.05 level. Since the Unix time stamps can only provide a duration between a peer rater’s
starting time to open a peer’s assignment and the submitting time of the peer assessment, we cannot conclude about the level of actual peer rater on-task engagement during this period. The study attempted to overcome such a limitation by using three minutes of the instructor time spent in rating as the cut off to identify the students who did not spend necessary time in peer rating.

Taking into account all the four indicators, we found that the extent of participation, course performance, rating time, and verified certificate registration together can only explain 0.8% of variance in converted values of distance index, which means that the overall model does not lead to practical significance, even if it were statistically significant.

5.2 Limitation, Implication and Future Research

For all of the four research hypotheses, the results of data analysis demonstrated positive linear relationships with the accuracy of peer assessment. This indicates that the directions of the sample data were consistent with our hypotheses. However, none of them had statistically significant result, and all the effect sizes are very small, showing little practical significance.

In this particular MOOC, there were limitations and restrictions in the type of the secondary data variables. So we were not able to have accurate measures of the independent variables. For example, weekly record from Coursera only indicated the number of lecture videos opened. More specifically, the video count number is not able to delineate on-task and off-task activity (e.g., shopping online, playing computer games,
and chatting on Facebook). Moreover, the video record is not a mutually exclusive number. For instance, a student with a record of two videos may have watched one of two lecture videos for that week twice, watched the two different videos for that week, or watched two videos from previous weeks. Also, for the rating time, we only know when a student opened a peer’s assignment and when he or she submitted the assessment result, but we do not know the exact time they really spent on grading the assignment. With these secondary data, we can only make a best guess about peer raters’ characteristics and learning activities in this MOOC. Therefore, in the future, we recommend to begin with measuring and investigating some controllable independent variables in a traditional online peer assessment first, and then apply the confirmed effective variables in a MOOC environment to collect more accurate and more meaningful data.

Another limitation is that the instructor’s scores, which have been used as a gold standard in this study, may also lack of accuracy. First, the instructor might make some randomly mistakes. Second, as a universally phenomenon, rater drift is difficult to avoid (Warshaw, Dyck, Allsworth, Stout & Keller, 2001). The instructor’s grading standard and criterion might change over time. Both random mistakes and systematic mistakes due to the instructor can impair the validity of peer assessment.

Since peer grades may be used as a main source of summative assessment, validity of peer assessment results should be an essential concern, and any additional methods to improve the validity of peer assessment is worthwhile to create a meaningful experience. One possible way to increase the validity of peer assessment in MOOC is to include the quality of peer assessment as a requirement to obtain verified certificate. Since MOOC students can use verified certificate as a proof of their accomplishment, it is
reasonable to assume that they would put more effort into tasks listed as a requirement of getting verified certificate than other voluntary tasks. In this MOOC, the four component variables of extent of participation are also parts of the requirements of obtaining certificate, but the quality of peer assessment is not. Consistent with our expectation, extent of participation of verified-certificate groups is significantly higher than non-verified certificate group, but this difference is not found in the validity of peer assessment. Therefore, to incentivize peer rater’s seriousness in making responsible judgment, course designer should include the accuracy of peer assessment as a requirement of earning a certificate.

Previous studies about traditional peer assessments have showed empirical evidence that training could improve the accuracy of peer assessment (Saito, 2008; Liou & Peng, 2009; Brand-Gruwel, van Merriënboer, & Sluijsmans, 2002; Xiao, 2011, Yurdabakan, 2012). However, we do not know whether this training effect is also applicable to a MOOC context. Coursera has a good training program for peer raters. However, in order to increase the overall completion rate in MOOCs, very fewer instructors require their students to pass or participate in the training program. Most of them have tried to encourage more students to complete the course by reducing requirements or lowering standards. It is possible that some MOOC students with good content knowledge as well strong will of making responsible judgment of peers’ assignments may fail to provide accurate rating result due to a lack of skill of grading and experience of peer assessment. Therefore, requiring peer raters to participate and passing the training program is another possible way to improve the validity of peer assessment in MOOC.
Researchers and MOOC designers ought to insist on the high standard of requirement for each MOOC rather than the completion rate. Regardless of the wide ranging intentions and goals of MOOC participants from those with pure curiosity to pursuit of certification for a job or education, MOOCs ought to provide meaningful learning experiences with implementation of reliable and valid peer assessment methods. Although there are strong criticisms and concerns regarding demanding tasks and low completion rates (Schulze, 2014; Cairns, 2013; Koller, NG, Do, & Chen, 2013), higher standards are favorable to increase the trustworthiness of MOOCs certificates and the quality of the MOOCs students. Moreover, accurate feedback is critical to meaningful learning. Admittedly, implementing peer assessment training or other methods to enhance the quality of peer assessment may lower the completion rate. At the same time, higher standards may attract more serious students and benefit the learners as well as MOOCs over time. On the other hand, if MOOCs raised the completion rate by reducing the evaluation standards or course content, the intentional dilution of academic standards will eventually lead to an undermining of the value of MOOC credentials which may depreciate the value of MOOCs compared to other available existing learning media such as TV, broadcast, online videos, and libraries.
REFERENCES

American Educational Research Association., American Psychological Association.,
National Council on Measurement in Education., & Joint Committee on Standards
educational and psychological testing. Washington, DC.

http://www.tonybates.ca/2014/10/12/what-is-a-mooc/.

Journal of Online Learning and Teaching, 10(1), 44.

Billington, H. L. (1997). Poster presentations and peer assessment: Novel forms of

Technology & Society, 12(4), 257-268.

doi:10.1080/0260293022000009311


from: http://www.infotoday.com/


