International Review of Research in Open and Distributed Learning

Peer Assessment for Massive Open Online Courses (MOOCs)

Hoi K. Suen

Volume 15, Number 3, July 2014

URI: https://id.erudit.org/iderudit/1065374ar
DOI: https://doi.org/10.19173/irrodl.v15i3.1680

See table of contents

Publisher(s)

Athabasca University Press (AU Press)

ISSN

1492-3831 (digital)

Explore this journal

Cite this note

Suen, H. (2014). Peer Assessment for Massive Open Online Courses (MOOCs). *International Review of Research in Open and Distributed Learning*, *15*(3), 312–327. https://doi.org/10.19173/irrodl.v15i3.1680

Article abstract

The teach-learn-assess cycle in education is broken in a typical massive open online course (MOOC). Without formative assessment and feedback, MOOCs amount to information dump or broadcasting shows, not educational experiences. A number of remedies have been attempted to bring formative assessment back into MOOCs, each with its own limits and problems. The most widely applicable approach for all MOOCs to date is to use peer assessment to provide the necessary feedback. However, unmoderated peer assessment results suffer from a lack of credibility. Several methods are available today to improve on the accuracy of peer assessment results. Some combination of these methods may be necessary to make peer assessment results sufficiently accurate to be useful for formative assessment. Such results can also help to facilitate peer learning, online discussion forums, and may possibly augment summative evaluation for credentialing.

IRRODL
THE INTERNATIONAL REVIEW OF RESEARCH IN OPEN AND DISTANCE LEARNING

Peer Assessment for Massive Open Online Courses (MOOCs)

Hoi K. Suen
Pennsylvania State University, U.S.A.

Abstract

The teach-learn-assess cycle in education is broken in a typical massive open online course (MOOC). Without formative assessment and feedback, MOOCs amount to information dump or broadcasting shows, not educational experiences. A number of remedies have been attempted to bring formative assessment back into MOOCs, each with its own limits and problems. The most widely applicable approach for all MOOCs to date is to use peer assessment to provide the necessary feedback. However, unmoderated peer assessment results suffer from a lack of credibility. Several methods are available today to improve on the accuracy of peer assessment results. Some combination of these methods may be necessary to make peer assessment results sufficiently accurate to be useful for formative assessment. Such results can also help to facilitate peer learning, online discussion forums, and may possibly augment summative evaluation for credentialing.

Keywords: Massive open online courses (MOOCs); peer assessment; formative evaluation; calibrated peer review; credibility index

In the past several years, massive spen online courses or MOOCs have erupted throughout the higher education landscape worldwide. These are typically audio, video, and textual instructional modules delivered via the internet and are free of charge. Enrollments in these courses have ranged from thousands to hundreds of thousands, typically from all around the world with about 1/3 of the enrollees coming from the United States and India (Waldrop, 2013). Numerous universities have developed and offered MOOCs on a trial basis. Hybrid degree programs that include a combination of traditional and MOOC courses, such as the master's degree in computer science program at the Georgia Institute of Technology in the United States, have emerged. Some companies that offer MOOC online platforms are attempting to license MOOC contents to be coupled with traditional in-class discussions and exercises provided by supportive instructional staff from traditional brick-and-mortar universities and offer the combination as blended courses. Many MOOCs are offering participants various recognitions for participation and completion, ranging from certificates of completion to online badges, to college credits. State legislators in the United States, such as those in California, are demanding universities accept MOOCs for credit. These rapid developments have given MOOCs the appearance of potentially replacing at least some of the traditional university resident instruction courses as well as some online courses.

Many observers have questioned whether these MOOCs can actually replace traditional brick-and-mortar instruction or even established online courses (e.g., Kauza, 2014). Others suggest that MOOCs 'cheapen' higher education and threaten the survival of high quality programs. Proponents, on the other hand, have complained that we might be asking MOOCs to meet a higher standard of quality than traditional instruction. Additionally, they cite the rising cost of tuition for higher education, coupled with the decreasing average annual income of families, as an unsustainable model that MOOCs might help address (e.g., Barber, Donnelly, & Rizvi, 2013).

Regardless of one's position, the general vision regarding MOOCs is that they constitute individual stand-alone, completely functional units that not only serve as another open educational resource, but can in fact lead to massive open learning.

From the Few to the Masses

It might be useful to pinpoint key differences between MOOCs and traditional university instructional modes, including those of large lecture classes as well as non-MOOC online courses. Perhaps the most obvious and also most critical difference between MOOCs and traditional classes is scale. While the largest of the traditional classes – large university classes in lecture halls – may have over 1,000 students, MOOCs typically have tens of thousands to over 100,000 students.

There may well be many social, economic, technological reasons for the emergence of MOOCs at this juncture. However, from a broad historical perspective, the development of MOOCs is a logical continuation of a trend in education, made possible currently by developments in communication technology and the internet. Historically, education as

a social institution has moved in a single direction: from the education of the privileged few to the education of the masses. This is particularly the case for higher learning and technical training. We started from the few teachers with few disciples (e.g., Socrates, Confucius, Shakyamuni, instruction in the monastaries of Taxila in India) approach; evolved next to a system of many teachers each having very few students (e.g., masters and apprentices; tutors/imams/zen masters and students); to formal educational institutions for aristocrats and the privileged (e.g., European universities, U.S. Ivy League colleges, Guozijian [国子监] in China, madrasas [مدرسة] in Muslim countries); to finally mass compulsory basic education and mass higher education (e.g., land grant colleges, GI Bill in the U.S.) with many teachers each teaching many students. The next logical step in this evolution may very well be universal open education for either self-actualization or credentialing. MOOCs promise to be a part of this next step in education.

From Teaching to Broadcasting

As we move from education of the privileged few to education of the masses, the learner-to-teacher ratio is increased at every stage. Access to the teacher by students is reduced and the learning experience is correspondingly diluted. A most important loss is the reduction in the opportunity to interact with teachers. While many aspects of the teaching/learning experience can be approximated through technology, the opportunity to interact with the teacher is an inverse function of the learner-teacher ratio and cannot be approximated without cloning the teacher. Current large university lecture classes attempt to approximate this cloning via the use of teaching assistants. MOOCs are taking this learner-to-teacher ratio problem to yet another level.

What is so important about student-teacher interaction? This has to do with what constitutes a sound educational process. The process involves a 3-step cycle: teach-learn-assess/feedback (cf. Frederiksen & Collins, 1989). The formative assessment or feedback step is critical to guide subsequent instruction and to ensure learning. In the earliest Socratic/Confucius mode, feedback from and to each individual student occurs naturally and is constant and continuous. As we moved over the ages through the apprentice to mass education modes, feedback to each individual student has become more and more sparse. Attempts have been made by some to put individual feedback back into the cycle by designing what is known as dynamic assessment (cf., Feuerstein, Feuerstein, Falik, & Rand, 2000; Haywood & Lidz, 2007). However, what that does is attempt to force into the system of mass education the original few-teacher, few-student mode and has not been practical. Hence, this approach is found primarily in special education, where individualized educational plans are often used or even legally mandated. It is also found in second language learning to some degree (e.g., Lantolf & Poehner, 2011). Many large lecture classes in universities also attempt to put back some small degree of individual feedback by breaking the class up into smaller 'recitation' sessions with teaching assistants or tutors.

Feedback and assessment in open and distance learning are inherently difficult to begin with (Chaudhary & Dey, 2013; Letseka & Pitsoe, 2013; Suen & Parkes, 1996). The problem of the reduction of individual feedback from, and interactions with, instructors becomes extreme in MOOCs. Due to the scale of MOOCs, feedback to individual students from the instructor has become virtually impossible. Yet, teaching without assessing whether the student has learned and without giving students feedback as to whether they have indeed learned the material correctly amounts to a one-way information dump or broadcasting, not education. A MOOC, in that form, would be essentially not different from the thousands of free how-to Youtube videos on the internet or the various free instructional videos provided by the Khan Academy (http://khanacademy.org) and cannot be considered a complete teaching-learning experience.

Attempts at Remedies

The teach-learn-assess cycle is essentially broken in a MOOC. Various attempts have been or are being made to re-introduce some degree of formative assessment feedback into the process to prevent it from becoming a one-way information dump or broadcasting show.

Many methods are suitable for feedback in an open distance learning environment in general. These include (a) automated tutors; (b) peer feedback; (c) auto-scoring of assignments; (d) reflective networks; (e) written comments; (f) oral comments; (g) meta-verbal; (h) emoticons; (i) self-checks; and (j) ePortfolio (Costello & Crane, 2013). Additionally, many developments in ICT have enabled feedback and assessment activities analogous to those of feedback activities in a traditional classroom. However, only a limited subset of these methods and technology are applicable to the scale of MOOCs.

In terms of assessment, some MOOCs offer online multiple-choice quizzes that are machine-scored as progress checks and feedback to students. At the end of each instructional module, a number of multiple-choice questions would be posed to the student. These questions are intended to gauge the student's mastery of the concepts and other contents covered in that module. The scores on these tests would indicate whether the student has adequately learned the material and the scores are given to the student as feedback. Students who do not do well would be encouraged to return to the previous module to review the materials. This approach is basically an online version of the old programmed-learning approach (Bloom, 1971; Skinner, 1968), popular briefly in the 1960s and 70s and quite limited in applicability since it is appropriate only for certain types of course contents where abilities to recall or to differentiate concepts or to interpret or extract information from text or graphics related to the subject matter are the only important instructional objectives. It is also challenging to most instructors to develop good quality multiple choice test items to measure high-level cognition such as analysis, synthesis, and evaluation, in Bloom's taxonomy. It is not appropriate for courses in which the desired evidence of learning is to have students demonstrate an

ability to generate ideas or produce a product, such as answer open-ended questions, write an essay, submit a report, design an artifact, engineer a process, or solve an ill-defined complex problem.

For open-ended writing assignments, automated essay scoring algorithms can be used (Balfour, 2013). These essay scoring programs have become more and more sophisticated and can detect many types of error in writing and can provide automated feedback to inform students of errors. An example of such an algorithm is the e-rater system used by the Educational Testing Services in the United States to score essays in the SAT test (see http://www.ets.org/research/topics/as_nlp). However, these programs are appropriate only when English writing ability is the construct of interest and are therefore appropriate only for MOOCs that teach English writing skills. Additionally, even when the objective of the course is writing ability, these programs can only detect errors in the more mechanical aspects of writing such as verb-noun agreement, run-on sentences and other grammatical or syntactical errors, and even organization to some extent, but are generally not capable of evaluating such abstract qualities as theme, humor, irony, coherence, and so on (Williamson, Xi, & Breyer, 2012; Zhang, 2013).

To provide feedback to students in general, in some cases instructors would provide answers to a limited number of most popular questions posted in the MOOC online discussion forum. The popularity of each question is often determined by a system of like/dislike votes similar to that used in Facebook. This, of course, is quite far from providing individual formative feedback and leaves the overwhelming majority of student questions unanswered. For the majority of the students, formative assessment and feedback would still be missing.

One solution that has emerged to address both the problem of the lack of formative feedback and that of a lack of revenue stream for the investment of resources in the development of MOOCs is to place MOOCs within a blended learning or flipped learning structure. This approach would have students view contents of a MOOC on their own and at their own pace. After learning the materials via the MOOC, they would attend local brick-and-mortar classes in which they would do assignments and participate in discussions with local instructors. While the MOOC portion may be free, the face-to-face sessions would be fee-based. Georgia Institute of Technology in the United States, for instance, has initiated a Master of Computer Science degree program for $6,000 to combine MOOCs with a large number of instructional tutors in a blended manner. Coursera is also attempting to license contents of existing MOOCs to be coupled with local instructional staff for credit-bearing courses at traditional universities.

This blended- or flipped-learning approach appears to be a workable alternative that would solve the central problems of assessment, feedback, and revenue. The flipped or blended learning mode is fundamentally quite similar to many advanced seminars in universities in which students are assigned take-home readings from textbooks or reference materials and are then to provide reports and participate in instructor-led

discussions in class. This approach to the use of MOOCs would fundamentally change its nature and function from its original promise of offering massive free universal education to those of a free multimedia, interactive analog of a traditional textbook (see e.g., Krause, 2014).

Finally, the single approach that is widely applicable to most, if not all, MOOCs is to use peer assessment and peer discussion forums to provide formative feedback to students. In this approach, fellow students within a MOOC are asked to evaluate student assignments and to provide feedback to other students. Unlike the use of multiple-choice quizzes or automatic essay scoring, it is applicable to all contents and assignments. It is also the most economical approach without the need to hire a large pool of support instructional tutors as in the case of blended learning models. It allows a MOOC to be a complete stand-alone educational tool without reducing the role of the MOOC to that of a multimedia interactive textbook.

Peer Assessment and Issues

There is a large body of literature about various aspects and effective practices of peer assessment in traditional classroom instruction (see Falchikov & Goldfinch, 2000; Gielen, Dochy, Onghen, Struyven, & Smeets, 2011; Li, Xiong, Zang, Kornbaher, Lyu, Chung, & Suen, 2014; Norton, 1992; Topping, 2005). In traditional classroom instruction, peer assessment has been commonly used to facilitate class discussions, often in small groups or dyads, often under the supervision and guidance of the teacher, and augmented by instructor assessment (Gielen et al., 2011). Peer assessment in MOOCs, however, exists in a very different environment. First, and most obviously, is the issue of scale. For a single assignment within a single MOOC, there are tens of, to over a hundred, thousand potential peer raters evaluating up to over a hundred thousand submissions. The logistics of linking raters and assignments are considerably more complex (Balfour, 2013). The second difference is that, because of the scale, there is little to no instructor mediation, supervision, or guidance. (Note that for flipped learning, the supervision exists in the traditional portion of the course, not within the MOOC.) A third difference is that MOOC participants are international. There is a large variation in native language, culture, value, and worldview among peer raters. Without a teacher overseeing the process, there is also little sense of obligation or incentive for students to take the peer assessment process seriously. It is, for example, known that MOOCs which employ peer assessments tend to have lower course completion rates (Jordan, 2013). It is not clear whether this low completion rate is an effect of the use of peer assessment or the result of asking students to submit open-ended assignment tasks instead of just clicking multiple choice answers. Such tasks also concomitantly necessitate the use of peer assessment.

Because of these differences, peer assessment in MOOCs will need to be 1) simple and easy to understand for students; 2) efficient in execution without occupying much time; and 3) limited in that each student rater is asked to rate no more than a handful of other students' assignments. In other words, peer assessment methods in MOOCs need to be

as scalable as MOOCs. These limitations would in turn lead to each assignment being rated by no more than a handful of peers realistically. The resulting assessment score data would then be one of a nested design with missing data in most cells. With a large enrollment for the course but only a handful of different peer raters per assignment, the distribution of rater abilities and knowledge for each assignment and between assignments will necessarily be uneven and imbalanced. Some assignments would be rated by excellent and knowledgeable raters while some would be rated by poor and uninspired raters.

In its most basic form, the process of peer assessment within a MOOC would be as follows: A scoring rubric is developed for an assignment, the latter usually in the form of a project, an artifact, or a written report, within an instructional unit in a MOOC. Students are instructed to complete the assigned project and submit it online. Each project is distributed to several randomly selected fellow students by asking the fellow students to view the project online. Each fellow student rater is then to rate the quality of the project based on the predetermined scoring rubric. Raters are also asked to provide some written comments. The mean or median rating score is taken as the score for the project. The score as well as the written comments are then made available to the original student who submitted the project. Through this process, each project is rated by no more than a handful of peer raters and each peer rater would rate no more than a handful of projects.

Accuracy of Peer Assessment Results and Remedies

Perhaps the most glaring problem with peer assessment is how trustworthy the results are. After all, within peer assessment, the performance of a novice is being judged by other novices. Is it possible that peer raters misjudge the quality of the submission even with the guidance of the predetermined scoring rubric? Is it possible that peer raters judge a submission highly because the raters and the submitter share the same set of common but erroneous misconceptions? Or equally troubling, is it possible that a peer rater judges a submission as poor due to the rater's own misconceptions about the subject matter? Without the mediation of an instructor, can erroneous peer assessment results actually harm learning? In spite of a few studies suggesting peer assessment results correlate well with instructor ratings in conventional classrooms as well as online courses for highly structured tasks with narrowly defined correct responses (e.g., Bouzidi & Jaillet, 2009), the doubt regarding the accuracy of peer assessment in general remains. Students, in particular, do not trust the results of peer assessment (e.g., Furman & Robinson, 2003). A similar problem exists for unmoderated peer online discussion forums.

To provide a glimpse of students' mistrust of peer assessment results or peer online discussions, below is a sample of comments from peer evaluators found in several MOOCs offered by the Pennsylvania State University in the U.S. in 2013 (Suen & Pursel, 2014):

> I hated the peer assessments as in some cases, their anonymity gave the peers an excuse to say mean-spirited things.
>
> Peer-to-peer evaluation can not replace the teaching by an expert. The evaluations are not deep and rich enough.
>
> Asking tens of thousands people to discuss online about anything is stupid. Letting three random Internet trolls (also known as peers) to decide whether one passes with distinction or not is moron.
>
> I really disliked the peer assessment. I worked very hard on my map and out of the reviews only one offered constructive criticism. The others I question if they even looked at my map rather than just the attached image of it. The comments that were made didn't even make sense.

A few approaches at various stages of development have been put forth to address the concern for accuracy of peer assessment results in MOOCs. These include connectivist MOOCs (cMOOCs), the Calibrated Peer Reviews (CPR™) system, a Bayesian post hoc statistical correction method, and a credibility index approach.

Connectivist MOOCs.

The approach used by connectivist MOOCs is to remove the concern for accuracy altogether from peer assessment and peer discussions by deliberately designing the course to welcome and encourage diverse perspectives from participants. Proponents of this approach view assignments, projects, and online discussions as opportunities for crowd-sourcing, leading to superior results that otherwise cannot be achieved individually by the students (or the instructor). The underlying orientation of this approach is what is known as the connectivist pedagogy, proposed by Siemens (2005) and others. The idea is that knowledge is gained experientially via connections a student makes among nodes. As such, peer perspectives provide the necessary nodes for the connections. The MOOCs with this basic orientation are referred to as cMOOCs. The idea of peer assessment is moot within a cMOOC paradigm, as peer connections are the very process of learning. This approach might be quite limited in terms of potentially applicable course contents. Further, the connectivist pedagogy remains controversial today (see, for example, Kirschner & van Merrienboer, 2013).

Calibrated Peer Reviews (CPR™).

Another approach is to evaluate the accuracy of the ratings provided by each student rater and assign weights to their ratings according to their relative degree of accuracy. The final rating score for the submission would be a weighted average of the rating

scores from peer raters. This approach is exemplified by the Calibrated Peer Review (CPR™) developed at the University of California – Los Angeles. The CPR™ approach is a general purpose peer assessment approach that is readily applicable to MOOCs. It is inherently scalable and can be used in MOOC as well as non-MOOC settings. The general peer assessment process is similar to that of the basic peer assessment approach, with the addition of a calibration process. During calibration, each peer rater is to rate up to three standard essays or projects of known quality that had already been rated by the instructor. All peer raters would rate the same essays/projects. The proximity between a peer rater's rating score and that of the instructor of the same essay/project is used as an indicator of the accuracy of the peer rater. This indicator is then used as the weight for that rater's ratings of actual peer performances. The more accurate is the rater, the more weight is given to that rater's judgment of peer performance. The performance score for each student submission is the weighted average of peer judgment scores. Many studies have demonstrated that CPR™ is an effective instructional tool that can help to improve students' scientific writing skills, confidence in self-assessment, academic performance in physiology, patient note writing among medical students, and so on (e.g., Furman & Robinson, 2003; Hartberg, Guernsel, Simpson, & Balaster, 2008; Likkel, 2012; McCarty, Parkes, Anderson, Mines, Skipper, & Grebosky, 2005; Pelaez, 2002; Reynolds & Moskovitz, 2008). However, few studies have been conducted to demonstrate the system can produce reliable and valid assessment results.

Bayesian post hoc stabilization.

Piech, Huang, Chen, Do, Ng, and Koller (2013), Goldin and Ashley (2012), and Goldin (2011) developed a number of Bayesian models to improve peer assessment results by imposing standard prior distributions to the ratings. The process proposed by Goldin is fundamentally similar to an empirical Bayes estimation process by imposing a normal prior within-rater distribution of rating scores as well as a normal prior between-rater distribution of scores. This process would produce more stable posterior peer assessment results, but cannot actually correct systematic errors of judgment due to misconceptions. This approach is shown to produce peer ratings that are more accurate than those from the basic peer assessment approach. Goldin (2011) found that the Bayesian approach reduced error of predicting instructor rating by 19-30%. The Piech et al. method is slightly different, but follows the same basic logical orientation. Whereas the CPR™ approach, as well as Goldin's approach, define accuracy as proximity to instructor rating, Piech et al.'s approach defines accuracy as proximity to the mean or median of peer rater scores in either a unidimensional or multidimensional space.

Credibility index.

The credibility index approach (Suen, 2013a, 2013b; Xiong, Goins, Suen, Pun, & Zang, 2014; see http://tlt.psu.edu/2013/07/12/peer-assessment-in-moocs-the-credibility-index/) is an attempt to improve the accuracy of peer feedback by modifying and refining the CPR™ method. The basic premise of the credibility index approach is that

errors in peer assessment results arise from at least three sources: basic error of judgment due to insufficient knowledge (inaccuracy), random judgmental error due to idiosyncratic situational factors at the time of judgment (inconsistency), and inability to maintain a constant level of accuracy from context to context (intransferability). Whereas the CPR™ method considers only the inaccuracy of the peer rater, the credibility index (CI) approach takes into consideration the accuracy of the rater, the consistency of the rater, and the transferability of the level of accuracy between contexts and assignments. The approach attempts to garner the needed additional information without adding much more to the rater's burden beyond what is already gathered in the CPR™ method. Theoretically, this approach should improve the accuracy of peer assessment results and there is preliminary evidence that is supportive of that claim (Xiong et al., 2014). Additional research is currently underway to confirm its efficacy. If proved to be effective, the CI can also be used to rank peer answers and comments in online discussion forums based on the CI value of each responder, and thus is potentially capable of moving the system away from ranking comments based on popularity to one based on knowledge.

Nature of Peer Assessment Errors

If MOOCs are to be a complete educational experience, and not just a free multimedia version of traditional textbooks, the key seems to be whether there is a viable and scalable built-in formative assessment and feedback process. Among the various options available, peer assessment is the most widely applicable method to date. In spite of the many studies showing the efficacy of peer assessment in promoting learning, skepticism remains as to whether peer assessment results can be trusted.

One source of ambiguity in evaluating the accuracy of peer assessment results seems to be the problem of determining what constitutes the true score. Most studies that attempt to evaluate accuracy have used instructor rating as the absolute standard and the quality of peer rating is determined by how far it departs from instructor rating. However, Piech et al. (2013) offer a different argument:

> For our datasets, we believe that the discrepancy between staff grade and student consensus typically results from ambiguities in the rubric and elect to use the mean of the student consensus on a ground truth submission as the true grade.

In the case of Piech et al.'s situation, the ground truth submission was rated by hundreds of peer raters. Given the large number of peer raters, their decision to use the mean of student ratings as the 'true score' may be a manifestation of the trust in crowdsourcing.

While the majority of studies continue to consider proximity to instructor rating as the gold standard of accuracy, Piech et al.'s reasoning does reflect the complexity of the

situation. There are at least six types of discrepancies in a peer assessment situation. These include: A) the discrepancy between the rating given by a peer rater and rating by the instructor on the same piece of work; B) the random situational fluctuations of the ratings to that same piece of work given by that same peer rater under different conditions; C) the inconsistency of ratings given to other similar pieces of work with similar quality but may differ in context or style; D) the random discrepancies between different peer raters on the same piece of work using the same set of criteria or rubric; E) the systematic discrepancy between different raters on the same piece of work due to difference in rater competence or rater leniency/stringency; and F) the random situational fluctuations of the ratings to the same piece of work given by the same instructor under different conditions. The situation is analogous to a moving archer on horseback shooting at a moving target.

Rater training and a carefully constructed rubric can help reduce some of the errors from all sources. However, in addition to rater training and good rubrics, the different approaches to peer assessment discussed earlier can be viewed as attempts to tackle different combinations of these sources of error. The CPR™ is designed to minimize errors A and E in general, but the existence of other sources of error can render this effort ineffective for a given assessment. The Bayesian approach is designed to minimize error D. The CI approach is designed to minimize errors A, B, C, & E, but does require slightly more information from the rater than otherwise collected by other methods. No method has been developed to minimize error F, except for the desirable practice of developing clear rubrics. The cMOOC approach would not consider these to be errors at all, but part of the diversity of views upon which knowledge is to be gained.

It is theoretically possible to combine these approaches into a single most effective composite approach in which raters are calibrated after training via the credibility index approach and the resulting ratings are further refined via a Bayesian or empirical Bayes approach.

Finally, one remaining problem with peer assessment in MOOCs is the probability of an assignment being rated by all poor raters. This problem may be minimized if the peer rater distribution algorithm uses a stratified sampling process based on prior knowledge, or credibility index value, or performance as a peer rater in previous assignments, instead of the current random assignment process.

It should be noted that peer assessment, whether the results are accurate or not, is considered valuable as an instructional tool in its own right. Indeed, Topping (2005) folded peer assessment as part of a larger category of peer learning. However, accurate peer assessment results would further enhance this learning experience, as well as serve the purpose of assessment. Additionally, if can be made reasonably accurate, peer assessment results can be used for purposes beyond formative assessment. One such potential use is to facilitate online discussion forums by putting more weight on opinions of student raters whose judgments of peer performances are close to that of the instructor's. Another potential use is to use student raters' performance-as-raters to

supplement final summative evaluations of each student for the purpose of credentialing. The feasibility of the latter purpose, even if peer assessment results are made accurate, is not clear at this time – at least not clear in the United States. Based on the 2002 US Supreme Court unanimous ruling on the *Owasso Public Schools v. Falvo* case (2002), peer assessment as formative evaluation does not violate the 1974 U.S. law known as the Family Education Rights to Privacy Act (FERPA). The key basis of the judgment seems to be restricted to the idea that peer assessment results for formative purposes do not constitute part of the student's school record. At this time, whether peer assessment results can be used as part of a summative grade, including credentialing and certification, and still not violate FERPA is not clear.

References

Anderson, T. (2008). Towards a theory of online learning. In T. Anderson (Ed.), *The theory and practice of online learning* (pp. 91–119). Athabasca, Canada: Athabasca University Press. Retrieved from: http://cde.athabascau.ca/online_book/second_edition.html

Balfour, S. P. (2013). Assessing writing in MOOCs: Automated essay scoring and calibrated peer review. *Research & Practice in Assessment, 8*. Retrieved from http://www.rpajournal.com/assessing-writing-in-moocs-automated-essay-scoring-and-calibrated-peer-review/

Barber, M., Donnelly, K., & Rizvi, K. (2013). *An avalanche is coming : Higher education and the revolution ahead*. London, England : Institute for Public Policy Research.

Bloom, B. (1971). *Mastery learning*. New York: Holt, Rinehart, & Winston.

Bouzidi, L., & Jaillet, A. (2009). Can online peer assessment be trusted? *Educational Technology & Society*, *12*(4), 257–268.

Chaudhary, S.V.S., & Dey, N. (2013). Assessment in open and distance learning system (ODL): A challenge. *Open Praxis*, *5*(3), 207-216.

Costello, J., & Crane, D. (2013). Technologies for learner-centered feedback. *Open Praxis*, *5*(3), 217-225.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, *70*(3), 287-322

Feuerstein, R., Feuerstein, S., Falik, L., & Rand, Y. (2002). *Dynamic assessments of cognitive modifiability*. Jerusalem, Israel : ICELP Press.

Frederiksen, J.R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, *18*(9), 27-32.

Furman, B., & Robinson, W. (2003). Improving engineering report writing with Calibrated Peer Review™. In D. Budny (Ed.), *Proceedings of the 33rd Annual Frontiers in Education Conference*. Piscataway, NJ: IEEE Digital Library.

Gielen, S., Dochy, F., Onghena, P., Struyven, K., & Smeets, S. (2011). Goals of peer assessment and their associated quality concepts. *Studies in Higher Education*, *36*(6), 719-735.

Goldin, I. M. (2012). Accounting for peer reviewer bias with Bayesian models. *Proceedings of the Workshop on Intelligent Support for Learning Groups at*

the 11th International Conference on Intelligent Tutoring Systems. Chania, Greece.

Goldin, I. M., & Ashley, K. D. (2011). Peering inside peer review with Bayesian models. In G. Biswas, S. Bull, J. Kay, & A. Mitrović (Eds.), *Proceedings of 15th International Conference Artificial Intelligence in Education* (Vol. 6738, pp. 90–97), Springer.

Hartberg, Y., Guernsel, A. B. , Simpson, N. J. , & Balaster, V. (2008). Development of student writing in biochemistry using calibrated peer review. *Journal for the Scholarship of Teaching and Learning, 8*(1), 29-44.

Haywood, C. H. & Lidz, C. S. (2007). *Dynamic assessment in practice: Clinical and educational applications*. New York: Cambridge University Press

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

Kauza, J. (2014). MOOC assigned. In S.D. Krause & C. Lowe (Eds.), *More questions than answers : Scratching at the surface of MOOCs in higher education* (pp. 105-113). Anderson, S.C. : Parlor Press.

Kirschner, P.A., & van Merrienboer, J.J.G. (2013). Do learners really know best? Urban legends in education. *Educational Psychologist, 48*(3), 169-183.

Krause, S. D. (2014). MOOC assigned. In S.D. Krause & C. Lowe (Eds.), *Invasion of the MOOCs : The promises and perils of massive open online courses* (pp. 122-129). Anderson, S.C. : Parlor Press.

Lantolf, J. P., & Poehner, M. E. (2011). *Dynamic assessment in the foreign language classroom. A teacher's guide* (2nd ed). The Pennsylvania State University, Center for Advanced Language Proficiency Education and Research.

Letseka, M., & Pitsoe, V. (2013). Reflections on assessment in open distance learning (ODL): The case of the University of South Africa (UNISA). *Open Praxis*, *5*(3), 197–206.

Li, H.L., Xiong, Y., Zang, X.J., Kornbaher, M., Lyu, Y.S., Chung, K.S., & Suen, H.K. (2014, April). *Peer assessment in a digital age: A meta-analysis comparing peer and teacher ratings*. Presentation at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Likkel, L. (2012). Calibrated Peer Review™ essays increase student confidence in assessing their own writing. *Journal of College Science Teaching, 41*(3), 42-47.

McCarty, T., Parkes, M.V., Anderson, T. T., Mines, J., Skipper, B. J., & Grebosky, J. (2005). Improved patient notes from medical students during web-based teaching using faculty-Calibrated Peer Review and self-assessment. *Academic Medicine, 80*(10 suppl.), S67-70.

Norton, S.M. (1992). Peer assessments of performance and ability: An exploratory meta-analysis of statistical artifacts and contextual moderators. *Journal of Business and Psychology, 6*(3), 38-399.

Owasso Public Schools v. Falvo, No. 00-1073. Retrieved from http://laws.findlaw.com/us/000/00-1073.html

Pelaez, N. (2002). Problem-based writing with peer review improves academic performance in physiology. *Advances in Physiology Education, 26*, 174-184.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013) *Tuned models of peer assessment in MOOCs*. Palo Alto, CA: Stanford University. Retrieved from http://www.stanford.edu/~cpiech/bio/papers/tuningPeerGrading.pdf

Reynolds, J. A., & Moskovitz, C. (2008). Calibrated Peer Review™ assignments in science courses: Are they designed to promote critical thinking and writing skills? *Journal of College Science Teaching, 38*(2), 60-66.

Siemens, G. (2005). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology & Distance Learning, 2*(1). Retrieved from http://www.itdl.org/Journal/Jan_05/index.htm.

Skinner, B. F. (1968). *The technology of teaching*. New York: Appleton-Century-Crofts.

Suen, H. K. (2013a, October). *Role and current methods of peer assessment in massive open online courses (MOOCs)*. Presentation at the First International Workshop on Advanced Learning Sciences (IWALS), University Park, Pennsylvania, U.S.A.

Suen, H. K. (2013b, November). *Peer assessment in MOOCs*. Presentation at the Educause Learning Initiative, 2013 Online Fall Focus Group. http://www.educause.edu/events/eli-online-fall-focus-session-2013

Suen, H. K., & Parkes, J.T. (1996). Challenges and opportunities for student assessment in distance education. *DEOSNEWS, 6*(7), file no. 96-00007.

Suen, H. K., & Pursel, B.K. (2014, March). *Scalable formative assessment in Massive Open Online Courses (MOOCs)*. Presentation at the Teaching and Learning with Technology Symposium, University Park, Pennsylvania, U.S.A.

Topping, K. J. (2005): Trends in peer learning. *Educational Psychology: An International Journal of Experimental Educational Psychology*, *25*(6), 631-645.

Waldrop, M. M. (2013, March). Online learning : Campus 2.0 : Massive open online courses are transforming higher education – and providing fodder for scientific research. *Nature*, *495*, 160-163. Retreived from http://www.nature.com/news/online-learning-campus-2-0-1.12590.

Williamson, M., Xi, X., & Breyer, F.J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2-13.

Xiong, Y., Goins, D., Suen, H.K., Pun, W.H., & Zang, X. (2014, April). *A proposed credibility index (CI) in peer assessment.* Presentation at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

Zhang, B. (2013). Contrasting automated and human scoring of essays. *R&D Connections*, *No. 21*(March). Princeton, NJ: Educational Testing Services.