International Review of Research in Open and Distributed Learning

Evaluating the Validity and Applicability of Automated Essay Scoring in Two Massive Open Online Courses

Erin Dawna Reilly, Rose Eleanore Stafford, Kyle Marie Williams and Stephanie Brooks Corliss

Volume 15, Number 5, November 2014

Special Issue: Research into Massive Open Online Courses

URI: https://id.erudit.org/iderudit/1065537ar
DOI: https://doi.org/10.19173/irrodl.v15i5.1857

See table of contents

Publisher(s)

Athabasca University Press (AU Press)

ISSN

1492-3831 (digital)

Explore this journal

Cite this article

Reilly, E., Stafford, R., Williams, K. & Corliss, S. (2014). Evaluating the Validity and Applicability of Automated Essay Scoring in Two Massive Open Online Courses. *International Review of Research in Open and Distributed Learning*, *15*(5), 83–98. https://doi.org/10.19173/irrodl.v15i5.1857

Article abstract

The use of massive open online courses (MOOCs) to expand students' access to higher education has raised questions regarding the extent to which this course model can provide and assess authentic, higher level student learning. In response to this need, MOOC platforms have begun utilizing automated essay scoring (AES) systems that allow students to engage in critical writing and free-response activities. However, there is a lack of research investigating the validity of such systems in MOOCs. This research examined the effectiveness of an AES tool to score writing assignments in two MOOCs. Results indicated that some significant differences existed between Instructor grading, AES-Holistic scores, and AES-Rubric Total scores within two MOOC courses. However, use of the AES system may still be useful given instructors' assessment needs and intent. Findings from this research have implications for instructional technology administrators, educational designers, and instructors implementing AES learning activities in MOOC courses.

IRRODL THE INTERNATIONAL REVIEW OF RESEARCH IN OPEN AND DISTANCE LEARNING

Evaluating the Validity and Applicability of Automated Essay Scoring in Two Massive Open Online Courses

Erin Dawna Reilly, Rose Eleanore Stafford, Kyle Marie Williams, and Stephanie Brooks Corliss
University of Texas at Austin, United States

Abstract

The use of massive open online courses (MOOCs) to expand students' access to higher education has raised questions regarding the extent to which this course model can provide and assess authentic, higher level student learning. In response to this need, MOOC platforms have begun utilizing automated essay scoring (AES) systems that allow students to engage in critical writing and free-response activities. However, there is a lack of research investigating the validity of such systems in MOOCs. This research examined the effectiveness of an AES tool to score writing assignments in two MOOCs. Results indicated that some significant differences existed between Instructor grading, AES-Holistic scores, and AES-Rubric Total scores within two MOOC courses. However, use of the AES system may still be useful given instructors' assessment needs and intent. Findings from this research have implications for instructional technology administrators, educational designers, and instructors implementing AES learning activities in MOOC courses.

Keywords: Massive open online courses; assessment; automated essay scoring systems

Introduction

A massive open online course (MOOC) provides online course content delivered by professors from top universities to any individual who chooses to enroll in the course. The subject of MOOCs is currently one of the most hotly debated topics in higher education. Proponents suggest that MOOCs could render traditional brick-and-mortar universities obsolete, while opponents maintain that high attrition rates and limited quality measures make MOOCs a threat to effective learning (Watters, 2013). As MOOCs have become more widespread, with some institutions offering badges or accepting MOOCs for credit, assessment has moved to the front and center of the conversation (Sandeen, 2013). A major question remains: Can MOOCs provide and adequately assess authentic, higher level student learning experiences?

Currently most assessment in MOOCs is based on computer-scored multiple choice questions, formulaic problems with correct answers, logical proofs, computer code, and matching items, often with targeted feedback based on the responses given (Balfour, 2013). While this type of assessment works well in certain disciplines, others rely more on open-ended writing assessments for students to fully demonstrate their learning. Many MOOC environments provide tools for delivering open-ended writing assignments and either self- or peer-scoring with a rubric, but the quality of the scoring and feedback can vary greatly, possibly making it inappropriate for high-stakes assessment. Consequently, there is a need for valid and reliable automated scoring of open-ended written assessments in MOOCs.

Open-Ended Assessment in Online Learning

Open-ended assessments are commonly used to measure students' writing skills, conceptual understanding, and higher order thinking skills such as evaluating, analyzing, and problem solving. By forcing students to construct a response rather than choose from a list of possible answers, students are more fully able to demonstrate what they know and are able to do. Several studies have highlighted the importance of open-ended writing assignments in facilitating higher level thinking, allowing students to make connections and think clearly and critically about important issues (Kellogg & Raulerson, 2007). A study of multiple choice versus essay writing assessments of second year college students found that essay prompts were associated with deeper level learning approaches, while multiple choice formats were more often associated with surface-level learning (Scouller, 1998). Open-ended assessments provide students with more opportunities to apply their knowledge and skills to authentic contexts and to transfer knowledge, while timely scoring provides feedback to students that leads to increased achievement (Chung, Shel, & Kaiser, 2006; Vonderwell et al., 2007; Wolsey, 2008; Gikandi, Morrow, & Davis, 2011; Crisp & Ward, 2008). For these reasons, open-ended assessment items enable students to demonstrate their higher level learning in a much richer fashion than other types of machine-scored items.

The use of open-ended responses in online course environments has become standard practice. Peers, teaching assistants, or instructors often use electronic rubrics to score open-ended responses and provide feedback to students. Timely feedback is particularly important in an online environment because it can (1) help break down barriers that exist for students seeking clarification of information (Wolsey, 2008); (2) enable students to quickly revise misunderstandings; (3) encourage sustained student engagement (Tallent-Runnels et al., 2006); and (4) promote student satisfaction (Gikandi et al., 2011). While the tools exist to gather open-ended assessment data from students in online environments, the scoring and feedback mechanism has proven problematic when scaling to large numbers of students.

Open-Ended Automated Assessment in MOOCs

Incorporating open-ended assessments with valid and reliable scoring has the potential to transform the MOOC experience, especially in the liberal arts disciplines. Several MOOC platforms have begun utilizing assessment tools that allow students to engage in critical writing and free-response activities. However, the large student populations make it impossible for course instructors to score all open response items. Peer assessment functionality exists, but ways of holding reviewers accountable for quality scoring and feedback often do not. In addition, recent studies have emphasized the importance of automatic feedback for asynchronous distance learners who cannot wait for instructor-specific feedback (Farrús & Costa-jussà, 2013). For these reasons the MOOC platform, edX, is experimenting with an automated essay scoring (AES) system that can quickly score student written responses.

The New York Times announcement of the innovative nature of the edX AES scoring tool generated discussion on several educational blogs (for example, Mayfield, 2013; Tan, 2013) and in the higher education press (for example, Markoff, 2013). The edX AES system uses an innovative machine learning algorithm to model the characteristics of responses at different score points using an instructor-developed rubric and approximately 100 instructor-scored student responses, which is a smaller number of required instructor-graded calibration essays than many other AES systems (Dikli, 2006). While AES systems have been around for several years, there are mixed results about their effectiveness.

The first AES system, known as the Project Essay Grader (PEG), was developed in 1966 as a potential grading strategy to help relieve teachers of the burden of grading essays for large classes. While this system was accurate at predicting human scores and had a fairly simple scoring method, critics of this early system argued that it measured only surface-level features of writing and could be deceived by students into giving higher scores to longer essays (Dikli, 2006). The *e-rater* system used by the Educational Testing Service (ETS) has been the subject of many AES-related articles and is generally found to be consistently predictive of scores given by human graders (Burstein & Chodorow, 1999). However, studies conducted by Wang and Brown on the *e-rater* resulted in significant differences between machine graders and human graders (2007)

and a lack of significant correlations among machine and human graders (2008), giving academics cause for concern. In 2012, AES critic Les Perelman submitted an essay to the ETS *e-rater* system composed of real words written in a nonsensical and incoherent way, and received the highest possible score for it (Gregory, 2013).

While still in the developmental phases, almost no research has been conducted on the validity, perceptions, and instructor best-practices of the edX AES system. Although the tool was successfully piloted in a chemistry course where 80% of students believed their score was accurate (J. Akana, personal communication, August 21, 2013), additional research is needed to calibrate and determine the reliability of the scores produced in different contexts and with different types of learners. An additional area for research is the differential use of holistic versus trait/rubric grading through AES systems. Holistic scoring involves giving one score based on an overall assessment of an assignment, while rubric (also referred to as analytic) grading refers to assigning multiple scores based on several features of an assignment; for example, analytic components of an essay might be clarity, organization, grammar, and spelling (Burstein, Leacock, & Swartz, 2001). The edX system utilizes both methods, creating both rubric-level and holistic scores for student essays, but records the holistic score as the final essay grade.

Overall, there is a growing call for research investigating the capabilities of AES tools, how faculty and students view and utilize them, and how they might be best embedded in MOOCs to promote greater critical thinking and interaction with course content, and to be used for high-stakes assessment. To address this concern, data was collected from the first two MOOCs to utilize the edX AES system. In this study, we investigated the following research questions: To what extent is the current edX machine-graded assessment system (both holistic and rubric-total) valid, reliable and comparable to instructor grading? Additionally, do the AES-graded assignments (AES-Holistic and AES-Rubric total) correlate with non-essay assignment grades in the course?

Study One

Method

Study One included MOOC student samples from an edX Pharmacy course in fall 2013, with an enrollment of approximately 15,000 students. The current study utilized a causal-comparative design, a non-experimental research design which involves data collection and analyses that allow for group comparisons upon a particular variable of interest (Martella, Nelson, & Marchand-Martella, 1999). In this study, the researchers examined data from three groups; specifically, comparisons were made between the AES-Holistic graded score group, the AES-Rubric graded score group, and the instructor-graded score group. Additionally, correlational analyses were used to

investigate potential relationships among AES- and instructor-scores and patterns of grading. Both causal-comparative and correlational designs have been used in prior AES studies to compare AES and human-grading as well, and were incorporated to more fully explore relationships among both mean differences and grading patterns (Wang & Brown, 2007; Wang & Brown, 2008).

The essay assignment involved students reflecting on patient compliance with medication prescriptions, and asked students to write a short-answer response of about 5 to 7 sentences. The instructor then graded 100 essays to calibrate the AES system. The rubric for the assignment consisted of 4 different general sections (Understanding, Support, Organization, and Content), on a scale of 0 -2, with total scores ranging from 0 to 8. Approximately 1,090 students completed this assignment, and 206 of the AES-scored essays were randomly selected, de-identified and re-graded by the instructor who originally calibrated the AES system, using the same rubric used for AES calibration.

Results

Prior to analyses, we statistically and visually inspected the score distributions for the three rating systems to assess their normality. We determined that the scores were substantially deviant from a normal distribution, which was indicated by excessive levels of skewness (AES-Holistic = -1.35, AES-Rubric = -1.98, Instructor = -2.12) and kurtosis (AES-Holistic = 1.89, AES-Rubric = 3.78, Instructor = 4.59) and inspection of frequency distributions, boxplots, and Q-Q plots. The non-normality of the score distributions was likely due to the eight-point scale used in calculating total essay scores. Therefore, all analyses used were non-parametric. Multiple analyses were conducted in order to determine the nature of the relationship between the two AES scoring systems (AES-Holistic and AES-Rubric Total) and the instructor's grading.

Wilcoxon signed rank tests (non-parametric repeated measures t-tests) were used to compare the average scores of each of the three essay scorers. Results indicated that there was a significant difference between the Instructor's and AES-Holistic's grading (S = -5731, $p < .0001$), such that the instructor gave students an average of 1.27 more points on the essay than the AES-Holistic grader. However, the AES-Rubric Total and Instructor scores did not significantly differ (S = 479.5, $p < .054$), with the instructor on average scoring essays .24 points higher than AES-Rubric Total. The averages of the two AES grading systems were also compared. The AES-Rubric Total was an average of 1.02 points greater than the AES-Holistic Score, which was a significant difference (S = 5404, $p < .0001$).

Spearman correlations found that there were significant relationships between all three essay grades. The highest correlation was between the AES-Holistic and AES-Rubric Total (r_s = .70, $p < .01$), with moderate correlations between each of these with the Instructor score (r_s = .59 and .57, respectively, $p < .0001$). Ordinal logistic regressions were used to predict expected Instructor total based on the AES scores. AES-Holistic scores significantly predicted instructor scores, B = .65 ($e^{0.65}$ = 1.91, $p < .0001$),

indicating that for every point given by the AES-Holistic scorer, the odds of a one-point gain in the Instructor score increases by a factor of 1.91. Correspondingly, there is a .65 probability that the instructor will give a point for each point that the AES-Holistic scorer assigns. The AES-Rubric Total was also found to be a significant positive predictor of Instructor score, $B = .59$ ($e^{.59} = 1.81$, $p < .0001$), meaning that as the AES-Rubric Total increases one point, the odds of a one-point gain in Instructor score increases by a factor of 1.81.This results in there being a .64 probability that the instructor will give a point for each point that the AES-Rubric scorer gives.

Percent agreement between the AES and Instructor grades were calculated. The agreement between individual rubric scores assigned by the Instructor and AES were high, ranging from 73.89% to 79.31% (see Table 1). Agreement between AES-Rubric Total and Instructor-total was lower though still relatively high, 55.17%. The percentage agreement was lowest between the AES-Holistic and Instructor grade, 17.24%.

Table 1

Percent Agreement between Instructor and AES-Scores – Pharmacy Course

	Rubric 1	Rubric 2	Rubric 3	Rubric 4	Rubric total	Holistic score
Percent agreement	79.31	77.83	75.37	73.89	55.17	17.24

Weighted kappas were also calculated to test whether there were significant differences between adjacent agreement scores. The AES-Holistic and Instructor Total weighted kappa coefficient was significantly different ($\kappa = .22$, $Z = 6.85$, $p < .0001$), indicating that there are significant differences in the grading of the AES-Holistic and Instructor. Similar findings were found for the AES-Rubric Total and Instructor Total agreement ($\kappa = .37$, $Z = 7.86$, $p < .0001$).

Lastly, Spearman correlations were conducted to determine the association between the three AES-essay grading systems and other course grades. These grades included the average of all homework assignments not including the essay grade, and the average of lab assignments, which were short quizzes following lecture videos and reading passages. All correlations were moderately low. The average lab grade had the highest associations with essay grades, having equal correlations with the AES grades ($r_s = .25$, $p < .00001$) and being the least associated with the Instructor Total ($r_s = .14$, $p < .05$). The correlations between average homework grade excluding the essay grade was most highly correlated with AES-Rubric Total ($r_s = .24$, $p < .0001$), followed by the AES-Holistic ($r_s = .22$, $p < .0001$), and the least with the Instructor Total ($r_s = .19$, $p < .01$).

Discussion

Past research suggests that, although the demand for AES systems is increasing, there is no consensus on the ability of these systems to automatically grade student essays and consistently predict instructor/human grading (Dikli, 2006). Most generally, the results of this study extend previous research by investigating the use of AES-Holistic and AES-Rubric systems in MOOCs, and how comparable they are to one another, instructor grading, and non-essay course grades (Deane, Williams, Weng, & Trapani, 2013; Rich, Harrington, Kim, & West, 2008; Shermis, Koch, Page, Keith, & Harrington, 2002). For this course, percent agreement between individual rubric scores assigned by the Instructor and AES-Rubric scorer were high, suggesting that inter-rater reliability on specific rubric criteria was moderately high for the AES-Rubric grader, though not for the AES-Holistic grader. Though high adjacent-agreement statistics are often easier to achieve than exact-agreement (see Cizek & Page, 2003), these results are promising for the use of the AES-Rubric grader.

Additional findings also emphasize the difference between holistic and rubric-total AES grading. Results indicated that the AES-Holistic Total and AES-Rubric Total were most highly correlated, which is consistent with research suggesting that trait ratings and holistic ratings are often correlated (Lee, Gentile, & Kantor, 2008; Deane, Williams, Weng, & Trapani, 2013) and can be as good or better than the correlation of ratings between two human raters (Shermis et al., 2002). Our data further suggest that for Study One, both AES systems tended to give lower scores than the instructor, and that these differences were most dramatic between the Instructor and AES-Holistic score. Consequently, these results indicate that the AES and instructor's scores are significantly related, but that the instructor assigned significantly higher grades than either AES-scoring system. This parallels past studies that have found instructors to grade higher than AES systems due to a more nuanced grasp of content, metaphor, and other rhetorical devices (Byrne, Tang, Tranduc, & Tang, 2010). However, the AES-Rubric Total and Instructor scores did not significantly differ, further suggesting that this particular AES system might be most comparable to Instructor grading when utilizing an AES-Rubric total score, as opposed to an AES-Holistic score.

It is also important to note that, although automatically-scored essay grades and other indicators of course success were moderately correlated, course grades appeared to be the least correlated with the instructor's essay grading. This may be due to the tendency of the instructor-total grades to be higher than the other AES grades and a subsequent ceiling effect, which could lead to lower course-essay grade correlations. In other words, little variability exists at extreme ends of a scoring scale when there is not a sufficient range of scores provided, and in this case it is possible that the instructor unintentionally created a maximizing effect by assigning higher essay grades (Keeley, English, Irons, & Henslee, 2013). Additionally, critics of AES systems have argued that they are unable to accurately score higher level writing tasks that would reflect authentic college-level learning and ability (Condon, 2013; McCurry, 2010). Consequently, these findings suggest a need to investigate the pedagogical differences between these

different assessment types in MOOCs, and how they might differentially measure learning objectives within MOOC education.

Study Two

Method

Study Two sought to replicate the findings from Study One using an assignment with a more elaborate rubric and generally longer essay responses. Similar to Study One, this study also utilized a combined causal-comparative and correlational study design to investigate both mean differences and relationships among AES-Holistic graded scoring, the AES-Rubric scoring, and instructor scoring of student essays. Participants included MOOC students from a fall 2013 philosophy course with approximately 29,000 students enrolled. The essay assignment asked students to reflect on a historical event and apply course-concepts to their analysis, with no word limit for responses. The rubric for the assignment consisted of 7 different general sections (Intelligibility, Clarity, Understanding, Support, Depth, Interpretation, and Comparison), on a scale of 0 -3, with total scores ranging from 0 to 21. Students first self-assessed their written work using the rubric, and then submitted their assignments for AES-grading. Approximately 423 students completed this assignment, and 128 of the AES-scored student surveys were randomly selected, de-identified and re-graded by the instructor who originally calibrated the AES system, using the same rubric used for AES calibration.

Results

Prior to analyses, we statistically and visually inspected the score distributions for the three rating systems to assess their normality. The distributions had levels of skewness (AES-Holistic = -0.77, AES-Rubric = -0.76, Instructor = -0.54) and kurtosis (AES-Holistic = 0.83, AES-Rubric = 0.69, Instructor = -0.51) within the appropriate ranges to be considered normally distributed. Based on this finding and a visual analysis of histograms, boxplots, and Q-Q plots, we determined that the scores were approximately normally distributed and that parametric statistical procedures were appropriate for the series of analyses. Paired samples t-tests were conducted to determine whether there were differences between the mean AES essay grades and Instructor Total. The average difference between the AES-Holistic score and Instructor total was .36, and not statistically significantly different (t = 0.88, p = .38). However, the AES-Rubric Total was significantly different than the Instructor Total (t = 2.43, p < .05), with the AES-Rubric Total being an average of 1.02 points higher than the Instructor Total. The AES-Holistic and AES-Rubric Total essay scores were also significantly different (t = 3.73, p < .001), with the AES-Rubric Total being an average of .66 points higher than the AES-Holistic Total.

Pearson correlations were conducted to investigate the associations between the three essay grading systems. The AES-Rubric Total and AES-Holistic had the highest correlation, r = .88 (p < .001). The Instructor Total had almost identical correlations with the AES-Holistic (r = .62, p < .0001) and the AES-Rubric Total (r = .60, p < .0001). Linear regressions analyses revealed that the AES-Holistic score was a significant predictor of Instructor Total (B = 0.92, t(1) = 8.79, p < .0001). This shows that for every one point given by the AES-Holistic, the Instructor Total is expected to increase by .92 points. The AES-Rubric Total was also a significant predictor of Instructor Total (B = 0.85, t(1) = 8.45, p < .0001), with every point increase on the AES-Rubric Total reflecting a .85 point increase in the instructor given essay score.

Percent agreement between the AES and Instructor rubric scores and essay grades were calculated. The agreement between individual rubric scores given by the Instructor and the AES system ranged from 35.94% to 50.00%. The Instructor Total had 14.06% agreement with the AES-Rubric Total and 13.28% agreement with the AES-Holistic score (see Table 2). To look further into the agreement between AES and Instructor scores, weighted kappas were calculated. The AES-Holistic and Instructor Total weighted kappa was significant (κ = .37, Z = 7.93, p < .0001), indicating that they differed in terms of weighted-score agreement. Analyses indicated that the essay grading by the AES-Rubric Total and the instructor also significantly differed (κ = .40, Z = 8.35, p < .0001).

Table 2

Percent Agreement between Instructor and AES-Scores – Philosophy Course

	Rubric 1	Rubric 2	Rubric 3	Rubric 4	Rubric 5	Rubric 6	Rubric 7	Rubric total	Holistic score
Percent agreement	46.88	49.22	46.09	35.94	50.00	40.63	42.97	14.06	13.28

Pearson correlations were calculated to analyze the relationship between the essay grades and another significant student-assessment. As a measure of non-essay student achievement, analyses utilized the “Lecture Sequence” average, which was the average of quizzes given after each video lecture. Correlations between AES-Holistic essay-scores and the Lecture Sequence average were small and non-significant (r = .11, p = .07). The AES-Rubric Total was also not significantly related to Lecture Sequence average (r = .10, p = .10). Instructor Total was not significantly related to the Lecture Sequence average, with a correlation of r = -.04 (p = .69).

Discussion

These analyses reveal that the AES grading systems were significantly correlated with Instructor Totals, though the instructor tended to assign slightly lower essay grades than both AES graders. Additionally, there was no significant difference between the Instructor and AES-Holistic scores, and all three grading systems (Instructor, AES-Holistic, and AES-Rubric) were positively, highly, and significantly correlated. This aligns with previous research suggesting that AES systems are often highly correlated (Johnson, 1996; Kakkonen, Myller, Sutinen & Timonon, 2008; Shermis et al., 2002).

Our findings suggest that, although asignificant mean difference existed between Instructor and AI-Rubric scores, there was actually high convergent validity among the three grading systems. This result is comparable to past studies indicating that well-developed AES systems can often produce grades comparable to skilled human graders (e.g., Shermis, Burstein, Higgins, & Zechner, 2010). For example, a study on an ETS research initiative called "Cognitively-Based Assessments of, for, and as Learning" (CBAL) by Deane and colleagues (2013) noted that perhaps rubric and holistic grading is best used when dividing up the grading tasks appropriately by grader. For example, the aspects of writing assessment that most closely match between the AES system and human graders (such as basic structure, grammar, and spelling) can be left to computers, while the more intricate aspects of writing quality, argumentation, and effective analysis can be reserved for human grading. Additionally, AES-essay grades and non-essay assignment grades were not correlated, corresponding with research highlighting the idea that different assignment types may measure different constructs in student learning or course outcomes (Scouller, 1998).

Though these analyses were encouraging regarding system validity, percent agreement analyses suggested that there is a significant discrepancy in the pattern of grading by both the AES systems and instructors. Specifically, inter-rater reliability analyses suggested that, on specific rubric criteria, the AES-Rubric total and Instructor scores were quite low. Findings such as these highlight the importance of using multiple metrics of validity and reliability when examining AES systems (Yang et al., 2001; Dikli, 2006). In other words, as AES tools continue to evolve and improve, it may be necessary to support these tools with supplemental measures of writing proficiency and ability, particularly in regards to the learning objectives being assessed within the writing task.

Another possible reason for the discrepancy in grading pattern may be attributed to the essay length, which has been shown to be highly correlated with both holistic and rubric scoring (Lee, Gentile, & Kantor, 2008). For example, Lee and colleagues' (2008) study on the relationship between individual rubric criteria scores and holistic scoring suggest that statistically controlling for essay length may aid in the usefulness and interpretability of rubric scores in AES systems. Along with other researchers of AES tools (Lee, Gentile, & Kantor, 2008; Shermis et al., 2002), we suggest that exploring the relationship between essay length, human grading, and AES scoring (both holistic and rubric) would be useful for future applications of automatic grading systems.

Conclusion

A series of different quantitative analyses was chosen to address the research questions, using the appropriate statistical analyses to obtain information on mean differences, correlational informational, and percent agreement examinations of different graders (AES-Holistic, AES-Rubric, and Instructor). Due to the amount of data and the subjective nature of essay grading that is a point of contention between proponents and critics of AES systems (Wang & Brown, 2007; Valenti, Neri, & Cucchiarelli, 2003), this methodology and various analyses methods was considered appropriate for both studies.

Overall, as the two study assignments had different rubrics and content, it is not reasonable to directly compare Study One and Study Two research outcomes. Additionally, as seen in the respective studies' discussions, there is literature to support both similarities and differences among AES-holistic, AES-Rubric, and instructor grading patterns. When considered separately, the results from Study One and Study Two suggest that the edX AES tool may not be a completely accurate and reliable tool for measuring student success on the writing assignments presented in these two MOOCs when compared to instructor grading. However, additional analyses for both Study One and Study Two revealed potential strengths of the AES system, such that either the AES-Rubric Total or AES-Holistic Total tended to be within one to two points of instructor grades. Overall, these results indicate a need for further analyses investigating specific algorithm scoring patterns on different essay aspects and rubric criteria.

This research suggests that, although statistically significant differences existed between instructor- and AES-grading for Study One and Study Two, the actual scores were often quite close. Consequently, depending on the intent of individual instructors for their chosen assignment, these systems may be more acceptable as a formative, as opposed to summative, assessment of student learning, as suggested by Shermis and Burstein (2003) and noted by Mahana, Johns, and Apte (2012). However, given these results, it is likely that instructors would not want to utilize this technology for high-stakes testing until further research and development of the tool is completed.

Limitations

Several limitations and recommendations based on the present studies should be noted. Though we have made some tentative comparisons between Study One and Study Two findings, the essay assignments were quite different in scope and subject. Due to these uncontrolled differences, we cannot make strong claims regarding clear reasons that account for statistical differences. More research is needed to determine the types of assignments that are most relevant for this scoring tool (length, topic, number of rubric categories, range of rubric scores, etc.). For the sake of comparability, future research may examine courses from more similar disciplines, with more similar assignments and grading scales, and may be useful with the integration of qualitative analyses. This

research was ultimately limited by the number of courses using the AES tool in the fall of 2013, constraining the study to evaluate only two courses used in these studies. As such, though we sought to replicate findings between courses and assignments, we are not able to compare them directly. Overall, with the growing availability of MOOCs for certificates or course credit, researchers have called for clarification and validation of the assessments utilized for MOOC students (Liyanagunawardena, Williams, & Adams, 2013).

Further research is also needed to investigate instructor perceptions of AES systems, and their pedagogical benefits and challenges. Specifically, instructors in these studies noted some key issues with the AES system and calibration. For instance, instructors noted several instances of plagiarism, and were unable to assign zero scores to these essays without affecting the essay-calibration system. Perhaps most importantly, this research was conducted on a particular AES system utilized through the edX platform; consequently, results may not generalize to other AES MOOC systems currently being utilized, tested, and developed.

Despite study limitations, the current research highlights potentially helpful next-steps for the creation, integration, and use of AES systems in MOOCs. For instance, AES-developers may want to consider using these systems in conjunction with an anti-plagiarism tool to reduce inflated scoring by the AES system of plagiarized essays. Faculty may also be more willing to engage with AES systems that offer greater metrics and information on holistic versus rubric-scored systems, and how they correlate with instructor grading. Finally, the fact that AES and Instructor-essay grades did not correlate highly with grades on other course assessments raises questions about how learning is measured in a MOOC and which assessment types are best suited to measure achievement of learning outcomes. Future studies should also be conducted based on more similar assignments in related fields for direct comparability and grading studies, as well as incorporating qualitative research evaluating AI-assessment tools. There is a growing demand for authentic assessment of higher-level learning in MOOCs, and research addressing these key issues in AES-systems would contribute greatly to that increasing need in online learning.

Acknowledgement

This study was supported by the MOOC Research Initiative (MRI), funded by the Bill & Melinda Gates Foundation.

References

Balfour, S. P. (2013). Assessing writing in MOOCs: Automated essay scoring and calibrated peer review. *Research & Practice in Assessment, 8*(1), 40-48.

Burstein, J., & Chodorow, M. (1999, June). *Automated essay scoring for nonnative English speakers*. In *Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing*. College Park, MD.

Burstein, J., Leacock, C., & Swartz, R. (2001). *Automated evaluation of essays and short answers*. *Proceedings of the 5th CAA Conference,* Loughbrough: Loughborough University.

Byrne, R., Tang, M., Tranduc, J. & Tang, M. (2010). *Journal of Systemics, Cybernetics & Informatics, 8*(6), 30-35.

Chung, G., Shel, T., & Kaiser, W. (2006). An exploratory study of a novel online formative assessment and instructional tool to promote students' circuit problem solving. *The Journal of Technology, Learning, and Assessment*, 5(6), 4-25.

Cizek, G. J., & Page, B. A. (2003). The concept of reliability in the context of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 125–145). Mahwah, NJ: Lawrence Erlbaum Associates.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, *18*(1), 100-108.

Crisp, V., & Ward, C. (2008). The development of a formative scenario-based computer-assisted assessment tool in psychology for teachers: The PePCAA project. *Computers & Education, 50*(4), 1509-1526.

Deane, P., Williams, F., Weng, V., & Trapani, C. S. (2013). Automated essay scoring in innovative assessments of writing from sources. *The Journal of Writing Assessment, 6*(1), 40-56.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment, 5*(1), 1-36.

Farrús, M., & Costa-jussà, M.R. (2013). Automatic evaluation for e-learning using latent semantic analysis: A use case. *The International Review of Research in Open and Distance Learning, 14*(1).

Gikandi, J., Morrow, D, & Davis, N. (2011). Online formative assessment in higher education: A review of the literature. *Computers and Education*, *57*(4), 2333-2351.

Gregory, M. A. (2013, April 26). *Computer thinks you're dumb: Automated essay grading in the world of MOOCs*. [Weblog]. Retrieved from: https://theconversation.com/computer-thinks-youre-dumb-automated-essay-grading-in-the-world-of-moocs-13321

Johnson, V. E. (1996). On Bayesian analysis of multirater ordinal data: An application to automated essay grading. *Journal of the American Statistical Association*, *91*(433), 42-51.

Kakkonen, T., Myller, N., Sutinen, E., & Timonen, J. (2008). Comparison of dimension reduction methods for automated essay grading. *Educational Technology & Society*, *11*(3), 275-288.

Keeley, J. W., English, T., Irons, J., & Henslee, A. M. (2013). Investigating halo and ceiling effects in student evaluations of instruction. *Educational and Psychological Measurement*, *73*(3), 440-457.

Kellogg, R. T., & Raulerson III, B. A. (2007). Improving the writing skills of college students. *Psychonomic Bulletin & Review*, *14*(2), 237-242.

Lee, Y.-W., Gentile, C., & Kantor, R. (2008). Analytic scoring of TOEFL® CBT essays: Scores from humans and *e-rater®*. *Educational Testing Service Research Report Series*, 1-71.

Liyanagunawardena, T., Williams, S., & Adams, A. (2013) MOOCs: A systematic study of the published literature 2008-2012. *The International Review of Research in Open and Distance Learning*, *14*(3).

Markoff, J. (2013). Essay-grading software offers professors a break. *New York Times*. Retrieved from http://www.nytimes.com/2013/04/05/science/new-test-for-computers-grading-essays-at-college-level.html?pagewanted=all&_r=0

Martella, R.C., Nelson, R., & Marchand-Martella, N.E. (1999). *Research methods: Learning to become a critical research consumer*. Boston: Allyn and Bacon.

Mayfield, E. (2013, April 8). *Six ways the edX announcement gets automated essay grading wrong*. [Weblog]. Retrieved on July 3, 2013 from http://mfeldstein.com/si-ways-the-edx-announcement-gets-automated-essay-grading-wrong/

McCurry, D. (2010). Can machine scoring deal with broad and open writing tests as well as human readers? *Assessing Writing*, *15*(2), 118-129.

Rich, C. S., Harrington, H., Kim, J., & West, B. (2008). *Automated essay scoring in state formative and summative writing assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, New York City, NY.

Sandeen, C. (2013). Assessment's place in the new MOOC world. *Research & Practice in Assessment, 8*(1), 5-12.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education, 35*, 453-472.

Shermis, M. D., & Burstein, J. (2003). *Automated essay scoring: A cross disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw & N. S. Petersen (Eds.), *International Encyclopedia of Education* (Vol. 4, pp. 20-26). Oxford, UK: Elsevier.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement, 62*(5), 5-18.

Tallent-Runnels, M.K., Thomas, J.A., Lan, W.Y., Cooper, S., Ahern, T.C., Shaw, S.M., et al. (2006). Teaching courses online: A review of the research. *Review of Educational Research, 76*(1), 93-135.

Tan, X. J. (2013, April 9). *Grading software sparks among academia.* Retrieved from http://www.thebehrendbeacon.com/news/all/grading-software-sparks-among-academia

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education, 2*, 319-330.

Vonderwell, S., Liang, X., & Alderman, K. (2007). Asynchronous discussions and assessment in online learning. *Journal of Research on Technology in Education, 39*(3), 309-328.

Wang, J., & Brown, M.S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment, 6*(2), 1-29.

Wang, J., & Brown, M. S. (2008). Automated essay scoring versus human scoring: A correlational study. *Contemporary Issues in Technology and Teacher Education*, *8*(4), 310-325.

Watters, A. (2013). MOOC mania: Debunking the hype around massive open online courses. *The Digital Shift*. Retrieved from: http://www.thedigitalshift.com/2013/04/featured/got-mooc-massive-open-online-courses-are-poised-to-change-the-face-of-education/

Wolsey, T. (2008). Efficacy of instructor feedback on written work in an online program. *International Journal on ELearning*, *7*(2), 311-329.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2001, October). *A review of strategies for validating computer automated scoring*. Paper presented at the meeting of the Midwestern Educational Research Association, Chicago, IL.