Global Times Call for Global Measures: Investigating Automated Essay Scoring in Linguistically-Diverse MOOCs

Erin D. Reilly, Kyle M. Williams, Rose E. Stafford, Stephanie B. Corliss
Janet C. Walkow, Donna K. Kidwell
University of Texas at Austin

Abstract

This paper utilizes a case-study design to discuss global aspects of massive open online course (MOOC) assessment. Drawing from the literature on open-course models and linguistic gatekeeping in education, we position freeform assessment in MOOCs as both challenging and valuable, with an emphasis on current practices and student resources. We report on the findings from a linguistically-diverse pharmacy MOOC, taught by a native English speaker, which utilized an automated essay scoring (AES) assignment to engage students in the application of course content. Native English speakers performed better on the assignment overall, across both automated- and human-graders. Additionally, our results suggest that the use of an AES system may disadvantage non-native English speakers, with agreement between instructor and AES scoring being significantly lower for non-native English speakers. Survey responses also revealed that students often utilized online translators, though analyses showed that this did not detrimentally affect essay grades. Pedagogical and future assignment suggestions are then outlined, utilizing a multicultural-lens and acknowledging the possibility of certain assessments disadvantaging non-native English speakers within an English-based MOOC system.

Literature Review

MOOCS and the Promise of Open-Access Education

Massive open online courses, or MOOCs, have been controversial in the field of education, particularly higher education and educational research and assessment (Dolan, 2014; Watters, 2013). MOOCs are generally defined as large courses offered for free using open access materials and available to anyone with an internet connection. Although these courses do not typically offer credit from the providing institution, MOOC enrollees have the opportunity to earn completion certificates or "badges" (Liyanagunawardena, Adams, & Williams, 2013). Since 2011, millions of people around the globe have registered for hundreds of MOOCs delivered primarily through platforms such as edX, Coursera, and Udacity. An oft-reported goal for some MOOC providers is to allow access to educational materials for international learners who might not otherwise be able to take a course on a given subject due to distance, country, or socioeconomic status (Sandeen, 2013). Other online platforms have attempted using this model of "open access" to increase educational resources; however, there are still questions about the financial sustainability of such a movement (Yuan & Powell, 2013).

One of the major concerns of providing access to higher education in a global community is the need for academic literacy in English (Hamel, 2007). MOOCs—like many mediums for transmission of knowledge—are spreading institutional and educational content via the lingua franca of English through a global community of learners, with both negative and positive consequences (Kim, 2012). Given that the United States is still the primary purveyor and home for MOOCs, it is unsurprising that English is the predominant language for both the courses and technology developers for these platforms (Young, 2013). However, the current literature on distance education and online learning only occasionally considers potential difficulties for the international student learning experience due to linguistic issues.

Though MOOCs are typically globally accessible, they are not free of the cultural norms and potential Western biases that exist throughout global education. One of the most cited issues in global education is the historic (and expanding) role of English as a "gatekeeper" in education, employment, business opportunities, and promotion opportunities (Pennycook, 1999; Phillipson, 1992). Open-education researchers are beginning to call for a movement from "open access" to "open course" models that more thoughtfully incorporate local and global knowledge through student collaboration and participation in order to reduce the barriers for international students (Morgan & Carey, 2009). For instance, Liu, Liu, Lee, and Magjuka (2010), reporting on the impact of cultural differences for the learning experience of international students in an online environment, found that language is still one of the more predominate obstacles to ESL student participation online. The authors also note that a more culturally sensitive course design and instructor perspective may elicit higher levels of participation from international students, particularly for those who have difficulty with the English language.

Similarly, the goal of open access encounters, combined with the barrier of English language proficiency, suggests a need for greater understanding of global pedagogies. In an article detailing her experience as an international learner in online classes, Tan (2009) suggested that instructors integrate more cultural awareness and opportunities, such as the use of lecture videos, to enhance language proficiency. Additionally, Tan noted that detailed explanations of course expectations regarding grades and assignments in syllabi would be helpful in attracting international students who may be unsure of specific course expectations, a finding similarly noted by Liu and colleagues (2010). These factors lead to concerns regarding the consequences of this educational system; specifically, will MOOCs enhance and diversify student understandings of content, or privilege and homogenize the experiences of Western students and education?

Open-Ended Assessment and Non-Native English Speakers

For decades, the use of open-ended assessment for non-native English speakers has been considered both necessary and problematic within multiple academic disciplines (see Casannave & Hubbard, 1992). Open-ended responses and essays require learners to supply information as part of an assessment, as opposed to multiple-choice items that require learners to select a correct answer from a list of options. When thoughtfully constructed by an instructor, open-ended assessments evaluate higher-level learning and offer the opportunity for learners to receive detailed, personalized feedback (Attali & Burstein, 2006). Though these opportunities do not necessarily mean that students will use feedback, such assignments are meant to allow meaningful improvement in learning or performance (Furnborough & Truman, 2009).

In an effort to gain immediate feedback and assistance, students struggling with writing in a non-native language often utilize translation resources to improve their performance. For instance, educational researchers have suggested that the use of online translation systems is often utilized for immediate feedback when writing in a non-native language (Karabulut, Levelle, Li, & Suvorov, 2012). In fact, some researchers have suggested that increasing collaboration between MOOCs and universities from around the world has further pushed the boundary of online machine-translation technology to help non-native English speakers complete open-ended MOOC assignments (Clifford, Merschel, & Munné, 2013).

In terms of the impact of translational technologies, research has been inconsistent regarding whether this has a beneficial impact on course performance. For instance, Larson-Guenette (2013) found that learners' consistent use of online translation sites to check their work in another language was related to higher levels of active engagement with an assignment. However, the use of immediate open-ended feedback to enhance student motivation and self-regulation may be particularly important in the learning process, especially for distance language learners (e.g., Furnborough & Truman, 2009).

Automated Essay Scoring and Massive Open Online Classrooms

Due to the extremely large classes inherent in MOOCs, hand-scoring of open-ended assessments is often impracticable; consequently, some MOOC platforms have begun to integrate automated-essay scoring systems (AES) to grade open-ended assessments (Mayfield, 2013). Given that these AES tools are still in developmental phases, very little research has been conducted on the validity, perceptions, and best practices of AES systems embedded in MOOC platforms (Reilly, Stafford, Williams, & Corliss, 2014). Much remains to be learned about the specific scoring results of the tools. For instance, some critics of AES systems have argued that they are unable to accurately score higher-level writing tasks reflecting student performance expected at the college level (Condon, 2013; McCurry, 2010). Others note that AES systems do not accurately reflect scores that would have been given by an instructor, and may not process the nuances of writing in the way that a human grader can (Balfour, 2013). Additionally, the range of useful feedback for students engaging in AES assignments—a key part of the value of open-ended assessments—may also be a serious pedagogical issue.

There has also been a desire for a greater investigation of how AES learning activities can be utilized and structured in a way that does not disadvantage non-native English speakers (Chen & Cheng, 2008). One study performed through the Educational Testing Service found significant differences between the final human score and their e-rater system scores across language groups (Burstein & Chodorow, 1999). However, this difference did not significantly affect agreement between independent human raters, and the writing features examined by the e-rater were generalizable between native and non-native English speakers.

Another study by Guo (2009) on the Analytical Writing Assessment of the Graduate Management Admission Test (GMAT AWA) found that the AES tool provided by Intellimetric did not unfairly grade test-takers of different ethnicities, non-native English speakers, or students writing in a non-English language. Additionally, Dikli and Bleyle (2014) evaluated the use of the Criterion AES system within an ESL course reported that the AES system itself was useful to students attempting to refine their English writing skills and communication abilities. On the other hand, these AES systems are unique in that they have a long history of use and research on their validity and reliability, whereas newer AES systems such as those emerging in MOOCs have not been as thoroughly investigated or refined. Consequently, critics remain skeptical of the general ability of emerging MOOC AES systems to score student writing depending on different student language demographics and ability levels.

To contribute further to this line of inquiry, the following study utilized a single-course case study design to investigate the use of the edX AES system for a global, linguistically diverse audience. This research extends the work of authors such as Dikli and Bleyle (2014) by independently assessing an open-source AES system that allows for an instructor-created assignment and rubric. The edX AES system allows instructors to input a self-created rubric, which is then used by the instructor to grade 100 essays in order to train the machine-learning algorithm to assign rubric scores in the same manner. After the system is trained, the AES system assumes grading responsibilities for the remaining essays. For this study, the AES system gave back both rubric-category level and rubric-total scores. In the past, the edX AES rubric-system has shown adequate reliability and validity when compared to human graders, though the AES rubric-total tends to align best with shorter essays (Reilly, Stafford, Williams, & Corliss, 2014). Using information gathered from this course, including student grade data and survey responses, we sought to investigate the following questions.

Research Questions

1. Do self-reported levels of written and spoken English proficiency predict AES- and Instructor-essay scores?
2. Are non-native and native English speakers graded significantly differently on essays by the AES-grading system and human raters?
3. Are non-native English speakers graded significantly differently on non-essay assessments, when compared to native English speakers?
4. When controlling for English-language proficiency level, do students who use an online translator program receive higher scores than students who do not?

Method

Participants

Participants included MOOC student samples from an eight-week edX Pharmacy course, Take Your Medicine: The Impact of Drug Development. Each week consisted of two learning modules, which were lecture modules and videos with transcripts available in English. Each module was immediately followed by learning assessments that quizzed students on their module comprehension. In addition, the faculty developed a weekly learning lab for students where they applied what they learned that week by finding external online resources to answer the lab questions. In comparison to other assignments, labs particularly expanded beyond both the in-class content and traditional assessment tools by requiring students to research multiple-choice items through English-language pharmaceutical websites. This was an optional extra credit activity and the majority of the students participated.

Overall, 1,090 MOOC students completed the open-ended writing assignment. The mean age of students was 30.54 years and students were approximately equivalent regarding gender distribution (male = 48.5%, female = 51.5%). In terms of highest education level, 23% reported having a high-school degree, 32% a bachelor's degree, 24% a master's degree, 8% a doctoral degree, and 13% reported "other." The course was extremely diverse in terms of native language and English-proficiency level. In fact, a number of students in China set up satellite sites or small study groups, overseen by a course TA, in order to share and collaborate on translated class materials.

Of the students who completed the essay assignment, 35.26% identified English as their first language (EFL). Of the 64.74% English second-language (ESL) speakers, 85 distinct languages were reported, with Spanish (10.74%), Hindi (3.93%), and Portuguese (3.7%) as the top three non-English languages. Students also reported their proficiency (1 = not at all proficient to 5 = completely proficient) in both written English (M = 3.86) and spoken English (M = 3.87). Reported written and spoken proficiency scores were highly correlated ($r = .80$, $p < .001$). Post-course survey responses revealed that a goal for many students taking this MOOC was to gain proficiency in English (26%).

Procedure

The course essay assignment asked students to write a short-answer response of about 5 to 7 sentences reflecting on issues related to patient compliance with medical prescriptions. The students were then asked to answer five language-related questions regarding whether English was their first language (yes/no), what their first language was (write in), if they used an online translation program to write their essay (yes/no), and their level of proficiency with both written and spoken English. Additional course information was also collected, including course grades, post-lecture quiz grades, and lab assignments.

In total, 203 essays were randomly selected and de-identified for the purposes of this study. First, the original course instructor graded 100 essays within the edX platform to calibrate the AES system, which were not included in the study analyses. The instructor then graded an additional 203 essays according to the rubric used originally within the course. These additional essays were randomly selected and consisted of 65 EFL and 138 ESL speakers. The assignment utilized a rubric measuring four different areas: competence/understanding, support, organization, and content, with total scores ranging from 0 to 8 (see Table 1). When investigating differences across English proficiency levels, analyses were run using the total of rubric scores assigned by the AES and instructor.

Table 1. Essay assignment rubric categories and point allocation.

	Rubric Categories and Scoring			
	Rubric 1: Competence	**Rubric 2: Support**	**Rubric 3: Organization**	**Rubric 4: Content**
	Do you understand the potential health concerns and issues well?	**Do you provide evidence to support your arguments?**	**Is your essay clear and well organized?**	**Does your essay make appropriate reference to some of these keywords /phrases?**
Zero Points	Displays some serious misunderstandings of the issues it addresses; logic unclear, oversimplifies significantly, and gets some fundamental points wrong.	Little to no evidence provided to support idea.	Lack of organization makes underlying logic of argument difficult to follow.	No keywords or phrases used referring to relevant content.
One Point	Generally displays a good understanding of these healthcare issues, but on some finer points misunderstandings and confusions remain; gets the basics right, but oversimplifies a bit and misses some details.	Includes some supporting evidence but argument is not fully developed. Partially explores topic citing evidence from other sources.	Some organization makes underlying logic of argument possible to follow, but improvements can be made.	Makes appropriate reference to one or two keywords or phrases.
Two Points	Displays an excellent understanding of the issues it addresses, gets both the basics and details right. Shows an appreciation for logical structure and depth of issue.	Includes strong supporting arguments. Fully explores health-consequences using solid evidence.	Very organized and persuasive, explanations for argument are detailed and precise.	Makes appropriate reference to three or more keywords or phrases.

Results

The distributions of Instructor- and AES-scores were statistically and visually analyzed for normality. We found that the data substantially deviated from a normal distribution, as indicated by excessive levels of skewness (AES = -1.98, Instructor = -2.12) and kurtosis (AES = 3.78, Instructor = 4.59), as well as inspection of frequency distributions, boxplots, and Q-Q plots. Shapiro-Wilk tests also indicated that the score distributions significantly differed from normality (AES = 0.67, $p < 0.0001$; Instructor = 0.64, $p < 0.001$). This non-normality was likely due to the eight-point scale used in calculating total essay scores; therefore, we used only non-parametric analyses. Means and standard deviations for rubric level and total scores by grader and English-speaking category are provided in Table 2.

Multiple analyses were conducted in order to determine the nature of the relationship between the AES scoring system and the instructor's grading. Spearman correlations indicated that there were significant positive correlations between essay scores and reported English proficiency, such that greater levels of spoken English proficiency ($r_s(202) = .24$, $p < .001$; $r_s(202) = .28$, $p < .001$) and written English skills ($r_s(203) = .26$, $p < .001$; $r_s(203) = .27$, $p < .001$) predicted both higher AES-graded and instructor-graded essay scores, respectively. Data collected on students' self-reported English language proficiency (first language information, written ability, spoken ability, and use of an online-translator program) was utilized to investigate grading differences by AES and Instructor. Wilcoxon-Mann-Whitney tests were used as a non-parametric version of an independent samples t-test. Findings indicated that MOOC students reporting a non-English first language were scored significantly lower than EFL students by both the AES total ($z = 2.94$, $p < .01$) and Instructor total ($z = 2.97$, $p < .001$).

Table 2. Average scores assigned by the AES and Instructor for EFL and ESL students.

	AES Scores		**Instructor Scores**	
Variable	**EFL**	**ESL**	**EFL**	**ESL**
Rubric 1	1.86 (0.43)	1.70 (0.60)	1.92 (0.32)	1.84 (0.46)
Rubric 2	1.82 (0.46)	1.57 (0.68)	1.92 (0.27)	1.75 (0.53)
Rubric 3	1.82 (0.53)	1.67 (0.62)	1.89 (0.31)	1.71 (0.49)
Rubric 4	1.88 (0.38)	1.72 (0.56)	1.78 (0.45)	1.66 (0.56)
Total	7.37 (1.41)	6.67 (1.92)	7.52 (1.13)	6.96 (1.61)

*Note. EFL = English-First Language students (n = 65). ESL = English-Second Language students (n = 138).

Percent agreement between the Instructor and AES was calculated to describe the overlap in the scorers (AES and instructor) on each rubric point and the total score. These were calculated across all students and for EFL and ESL students separately (see Table 3). We also analyzed χ^2-statistics to determine whether the rate of Instructor-AES agreement differed by English-language category. Agreement on individual rubric scores ranges from 77% to 85% EFL students and from 72% to 78% for ESL students. Though percent agreement between Instructor and AES rubric-category scores was consistently higher for native English speakers than ESL speakers, these differences were not statistically significant. However, the percent agreement of total scores was significantly higher for EFL students than ESL students ($\chi^2 = 7.64$, $p < .01$), being separated by a margin of over 20% (EFL = 69%; ESL = 49%).

Table 3. Percent Agreement between Instructor- and AES-Scores

Score	Rubric 1	Rubric 2	Rubric 3	Rubric 4	Total
EFL	83.08%	84.62%	81.54%	76.92%	69.23%
ESL	77.54%	74.64%	72.46%	72.46%	48.55%
χ^2	0.83	2.55	1.96	0.46	7.64^{**}

Note. EFL = English-First Language students. ESL = English-Second Language students. χ^2 = test for association between English-language category and percent agreement, $^{**}p < .01$

As the percent agreement may have been higher for native English speakers due to the high abundance of perfect scores and lower variability in this group, we analyzed several other inter-rater agreement indices that account for the probability that the AES and instructor scores would agree due to the prevalence of scores assigned. The chance-corrected coefficients analyzed were Cohen's κ, Scott's π, and Krippendorf's α, which can all be interpreted as the percent agreement between the instructor and AES above that which is expected by chance. These coefficients produced consistent results and indicated that there was slightly greater instructor-AES agreement for EFL students than for ESL students, after accounting for the likelihood of ratings agreeing due to chance (see Table 4). However, the agreement rates of the English categories were very similar, and lower than would be considered acceptable in most social science applications, with the instructor and AES agreeing 21%-25% of the time after deducting the probability they would agree due to chance.

Table 4. Chance-corrected agreement between instructor and AES-Scores

	κ	π	α
Native English Speakers	0.24	0.24	0.25
Non-Native English Speakers	0.22	0.21	0.22

κ = Cohen's Kappa. π = Scott's Pi. α = Krippendorff's Alpha.

Wilcoxon signed rank tests (non-parametric repeated measures t-tests) were used to compare the average scores of students who did and did not use an online translator for essay writing, after matching students on English written proficiency and English-as-first-language status. Results suggested that, when controlling for English language proficiency, there was not a significant difference in AES total scores between students who used an online translation program versus those that did not (S = 309, p = .07). Wilcoxon-Mann-Whitney tests revealed that, for native and non-native English speakers, average post-lecture comprehension quiz scores (z = 1.12, p = .26) and average course grades (z = 1.45, p =.15) did not significantly differ, although lab grades (z =2.75, p <.01) were significantly higher for native English speakers.

Discussion

As MOOCs continue to serve a growing global audience, the need for linguistically sensitivity and globally applicable assessments of learning will also continue to grow. Our findings suggest that non-native English speakers are graded significantly lower by the AES grading system as well as by the instructor, when compared to native English speakers. Additionally, the results from this study show a positive and significant relationship between English proficiency and essay scores in total, with higher written and spoken English-language proficiency correlating with higher AES rubric scores and instructor scores.

This indicates that students who rate their English-language proficiency level more highly tend to receive better scores on their essays from the AES system as well as the instructor, while students who rate their proficiency level lower tend to receive lower essay scores from both graders. On the other hand, the level of score-agreement between instructor and AES grading was higher for native English speakers as compared to ESL students, even when reanalyzed to correct for chance agreement; this suggests that the AES system may in fact be less valid and comparable to human grading when scoring non-native English speakers.

Taken in conjunction, these findings suggest that although non-native English speakers performed lower on this assignment overall, the AES system itself may differentially disadvantage non-native English speaking students. Additionally, research suggests that open-ended assessment itself benefits native English speakers, as has been shown in previous literature on the use of essays in global distance learning (Goodfellow & Lea, 2005). Our finding that native English-speaking students performed significantly better on the lab exercises, which required online and English-based research on external websites, suggests that other types of assessment that draw on literacy-based skills may also disadvantage students with a non-English first language. Consequently, the necessity of multiple forms of assessment, as well as other low-stakes open-ended writing assignments (e.g., discussion boards), should also be assessed and utilized in order to support students in freeform learning experiences.

Finally, the findings from this study did not support the hypothesis that the use of an online translator would result in higher AES system scores after controlling for English-language proficiency level; in fact, there was no difference between students' rubric total scores in regards to their use of an online translation program. This suggests that the use of machine translation for non-native English speakers does not significantly impair, nor enhance essay quality, which supports research suggesting that web-based translators are not necessarily effective in translating text into another language in a way that aids in the quality of essay assignments (Williams, 2006). Further research in the area of AES systems and online translation programs may shed light on the strengths and shortcomings of using AES grading in non-native speaker populations as an assessment tool, given the increasing availability of free translation software.

Together, these results suggest that differences may exist between native and non-native English speakers when students are graded by AES systems, which is a clearly complex problem when examining the intentions of MOOC audiences. MOOCs have been hailed as an educational resource for learners outside of the United States to have quality access to educational resources and valuable learning experiences (Byerly, 2012; Meyer & Zhu, 2013). However, non-native English speaking populations appear to be at a disadvantage because of their language proficiency on certain assessments, and thus may not be well-served by MOOCs that intend to use essay or research-oriented assignments for high-stakes testing in their courses. MOOC issues must be examined in terms of equity and adequacy of global education, such that students might experience "linguistic gatekeeping" if such assessment types are used

as part of high-stakes testing in global courses (Phillipson, 1992). As Cushing Weigle (2013) notes in her study of AES systems and language diversity, the benefits of implementing a more expedient system for scoring writing should be weighed against the potential for marginalizing non-native English speakers within an English-based education system.

Multicultural and Pedagogical Considerations

In terms of fairness in learning and assessment, these findings have some prospective practical applications and future research suggestions. When the group of learners MOOCs are attempting to reach are also placed at a disadvantage for the purposes of evaluation and grading, educators must re-examine the usefulness and applicability of such assessments. For example, some educational researchers have suggested a global-education paradigm shift from emphasizing "native-like fluency" in English, and instead promoting "reasonable competence" (Methitham, 2009). Further research on the AES tool in other disciplines may help MOOC instructors and instructional designers better understand the ways in which this tool could be used to support student learning, as well as allow for refinement of the AES system algorithm to more accurately assess non-native English speakers' writing ability (Cushing Weigle, 2013; Dikli & Bleyle, 2014).

Culturally-sensitive research methodologies suggest the importance of taking language proficiency into account throughout the assessment and course-design process, in order to address potentially problematic assessments and grading issues (Uzuner, 2009). After engaging with the tool, it has been suggested that instructors hand-grade student answers after using the AES system to better determine potential rubric issues and levels of subject mastery (Walkow & Reilly, 2014). Ideally, students would respond in their native language, though this is difficult to accommodate when dozens of countries and languages are represented. One possible option would be identifying non-English language options (Spanish, Hindi, Arabic, etc.) for future AES systems to provide a more inclusive environment.

On the other hand, many students reported gaining greater English proficiency as a goal for this MOOC. Thus, creating formative opportunities for students to engage in English-based open-ended assignments might also be pedagogically useful. Such assignments in globally-available open courses may allow students to more actively achieve their own learning goals through self-regulation, beyond specific course outcomes (Peters, 2002). In order to improve access to course content globally—as opposed to "educating the educated" (Hollands & Tirthali, 2014, p. 13)—it is important that faculty, course developers, and platform programmers take into account the diversity of learners in MOOCs, and identify ways to help them meet their course goals.

Limitations

Several limitations for this study were present. First, language information was only acquired for students who completed the essay assignment. As Dikli (2006) has pointed out, AES systems that utilize prior calibration for grading accuracy can be only as good as what they learn from the initial calibration sample. Consequently, it is likely that many non-native English students opted out of this assignment, and thus did not have their data included in this analysis. Second, plagiarism was not accounted for by either the instructor or the AES system in this scoring set, though the instructor did find multiple plagiarized essays that were scored highly by the AES system. Future MOOC AES systems should attempt to incorporate plagiarism software in order to investigate and possibly ameliorate this issue. Additionally, investigating the usefulness of AES feedback in helping ESL students improve their writing might also help instructors to utilize such assignments for formative feedback. Third, our analysis was specifically

limited to one AES system within one course; additional testing should be conducted to examine the generalizability and applicability of our findings and suggestions.

Finally, methodological issues related to scale truncation might account for some of the significant differences found between instructor and AES scoring consistency. For the instructor-AES agreement analyses, coefficients are likely to be conservative in their agreement estimation due to the heavily skewed score distributions. In other words, these estimates may be lower than would be expected due to the penalties they assign for a large number of scores falling in a score category (Feinstein & Cicchetti, 1990). The majority of students in our sample received high scores and this ceiling effect is likely to have contributed to the great discrepancy between percent agreement and the chance-corrected agreement indices. Future research should be conducted using a rubric with more scale points, in order to better differentiate between student ability levels and better conduct comparability analyses.

Conclusion

Despite study limitations, our findings suggest that ESL students were consistently given lower scores between the AES and instructor grader. This finding suggest that future research may consider incorporating DIF analysis into AES systems in order to decipher whether these differences are a product of differential item functioning or impact. Though we were unable to conduct this analysis due to a low number of participants for the instructor-grading sample, we would encourage future researchers investigating language effects and AES use in MOOCs to utilize this methodology. Overall, our findings revealed that both the AES and instructor-graded non-native English speakers lower, and that instructor and AES score comparability was better for native English speakers. This research suggests the need to further evaluate the use of AES graders in MOOCs for non-native English speakers. These systems might be better utilized as formative, low-stakes assessments or to help students reach goals around English literacy. Additionally, future research may need to qualitatively evaluate the use of AES systems for non-native English speakers, in terms of usefulness of feedback, navigability, and perceptions of assignment fairness.

Acknowledgement

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

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning, and Assessment, 4*(3), 1-31.

Balfour, S. P. (2013). Assessing writing in MOOCs: Automated essay scoring and calibrated peer review™. *Research & Practice in Assessment, Special Issue: MOOCs & Technology, 8,* 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.

Byerly, A. (2012). Before you jump on the bandwagon…*Chronicle of Higher Education, 59*(2), 34.

Casanave, C. P., & Hubbard, P. (1992). The writing assignments and writing problems of doctoral students: Faculty perceptions, pedagogical issues, and needed research. *English for Specific Purposes, 11*(1), 33-49.

Chen, C-F. E., & Cheng, W-Y. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology, 12*(2), 94-112.

Clifford, J., Merschel, L., & Munné, J. (2013). Surveying the Landscape: What is the Role of Machine Translation in Language Learning?. @ tic.revista d'innovació educativa, 10, 108-121.

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

Cushing Weigle, S. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing, 18*, 85-99.

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

Dikli, S. & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback?. *Assessing Writing, 22*, 1-17.

Dolan, V. L. B. (2014). Massive online obsessive compulsion: What are they saying out there about the latest phenomenon in higher education? *The International Review of Research in Open and Distance Learning, 15*(2), 268-281.

Goodfellow, R. & Lea, M.R. (2005). Supporting writing for assessment in online learning. *Assessment & Evaluation in Higher Education, 30*(3), 261 – 271.

Guo, F. (2009). Fairness of automated essay scoring of GMAT® AWA. Graduate Management Admission Council® Research Report. Retrieved from http://www.gmac.com/~/media/files/gmac/research/research-report-series/rr0901_awafairness.pdf

Feinstein, A.R., & Cicchetti, D.V. (1990). High agreement but low kappa: The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543-549.

Furnborough, C., & Truman, M. (2009). How do adult beginner distance language learners respond to assignment feedback? A qualitative study. *Distance Education, 30*(3), 399-418.

Hamel, R.E. (2007). The dominance of English in the international scientific periodical literature and the future of language use in science. *AILA Review 20*, 53-71.

Hollands, F., & Tirthali, D. (2014). Why do Institutions Offer MOOCs?. *Online Learning: Official Journal of the Online Learning Consortium, 18*(3). Retrieved from http://olj.onlinelearningconsortium.org/index.php/jaln/article/view/464

Karabulut, A., Levelle, K., Li, J., & Suvorov, R. (2012). Technology for French learning: A mismatch between expectations and reality. *CALICO Journal, 29*(2), 341–366.

Kim, J. (2012, March). Why every university does not need a MOOC. *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/blogs/technology-and-learning/why-every-university-does-not-need-mooc

Larson-Guenette, J. (2013), "It's just reflex now": German language learners' use of online resources. *Die Unterrichtspraxis/Teaching German, 46,* 62–74.

Liyanagunawardena, T. R., Adams, A. A., & Williams, S. A. (2013). MOOCs: A systematic study of the published literature 2008-2012. *The International Review of Research in Open and Distance Learning, 14*(3), 202-227.

Liu, X., Liu, S., Lee, S.-h., & Magjuka, R. J. (2010). Cultural differences in online learning: International student perceptions. *Educational Technology & Society, 13*(3), 177-188.

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

Methitham, P. (2009). *An Exploration of culturally-based assumptions guiding ELT practice in Thailand, a non-colonized nation*. Doctoral dissertation, Indiana University of Pennsylvania.

Meyer, J. P. & Zhu, S. (2013). Fair and equitable measurement of student learning in MOOCs: An introduction to item response theory, scale linking, and score equating. *Research & Practice in Assessment Special Issue, 8*, 26-39.

Morgan, T., & Carey, S. (2009). From open content to open course models: increasing access and enabling global participation in higher education. *The International Review of Research in Open and Distance Learning, 10*(5).

Pennycook, A. (1999). Introduction: Critical approaches to TESOL. *TESOL Quarterly, 33*(3), 329-348.

Peters, O. (2002). Distance education in transition: New trends and challenges. Oldenburg: Bibliotheks- und Informations system der Universität Oldenburg.

Phillipson, R. (1992). *Linguistic imperialism.* Oxford: Oxford University Press.

Reilly, E. D., Stafford, R. E., Williams, K. M., & Corliss, S. B. (2014). Evaluating the validity and applicability of automated essay scoring in two massive open online courses. *The International Review of Research in Open and Distributed Learning*, *15*(5).

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

Tan, F. (2009). Tri-fold transformation: An international adult students' reflections on online learning. *Adult Learning, (20)*3-4, 38-40.

Uzuner, S. (2009). Questions of culture in distance learning: A research review. *The International Review of Research in Open and Distance Learning*, 10(3).

Yuan, L., & Powell, S. (2013). MOOC and open education: Implications for higher education. Centre for

Educational Technology & Interoperability Standards (white paper). Retrieved from http://publications.cetis.ac.uk/2013/667

Young, S. L (2013). MOOCS: Revolutionizing education or the latest business opportunity?. *Fukuoka University Review of Commercial Sciences, 58*(1-2), 173-187.

Walkow, J. C., & Reilly, E. (2014). Are we ready for robots to grade? *The Chronicle of Higher Education, 61*. Retrieved from http://chronicle.com/article/Are-We-Ready-for-Robots-to/148723/

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/

Williams, L. (2006). Web-Based Machine Translation as a Tool for Promoting Electronic Literacy and Language Awareness. *Foreign Language Annals, 39*: 565–578.