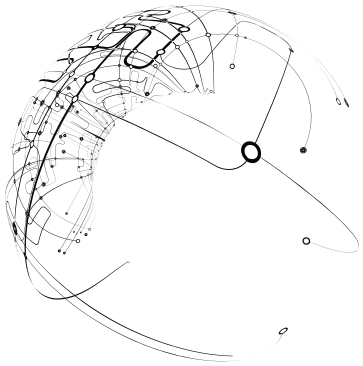

Abstract

Two of the largest Massive Open Online Course (MOOC) organizations have chosen different methods for the way they will score and provide feedback on essays students submit. EdX, MIT and Harvard's non-profit MOOC federation, recently announced that they will use a machine-based Automated Essay Scoring (AES) application to assess written work in their MOOCs. Coursera, a Stanford startup for MOOCs, has been skeptical of AES applications and therefore has held that it will use some form of human-based "calibrated peer review" to score and provide feedback on student writing. This essay reviews the relevant literature on AES and UCLA's Calibrated Peer Review™ (CPR) product at a high level, outlines the capabilities and limitations of both AES and CPR, and provides a table and framework for comparing these forms of assessment of student writing in MOOCs. Stephen Balfour is an instructional associate professor of psychology and the Director of Information Technology for the College of Liberal Arts at Texas A&M University.

AUTHOR

Stephen P. Balfour, Ph.D.
Texas A&M University

CORRESPONDENCE

Email
balfour@tamu.edu

ACKNOWLEDGEMENTS

Joshua Brown, Adam Mikeal, and Alysha Clark provided substantial feedback that greatly enhanced the value and clarity of the information in this article.

Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review™

Massive open online courses (MOOCs) allow any student to enroll in a course as long as they are able to access the course material online. MOOCs take advantage of various web-based technologies including video presentation, computer-based assessments, and online communication forums so that thousands of students can have access to all the course content, formative and summative assessments, and support from their fellow students. Some MOOCs have enrolled more than 150,000 students (DiSalvo, 2012); consequently, the time an instructor spends teaching and evaluating work per student is very low in high enrollment MOOCs. MOOCs use computers to score and provide feedback on student activities and assessment and thus rely heavily on multiple choice questions, formulaic problems with correct answers, logical proofs, computer code, and vocabulary activities. Scoring and providing feedback on written assignments in MOOCs has been the subject of a number of recent news articles.

Two of the largest MOOC organizations have announced mechanisms they will use to support the assessment of written work. EdX, MIT and Harvard's non-profit organization, has announced that it will use automated essay scoring (Markoff, 2013). They plan to use an application developed by a team including Vik Paruchuri who, with Justin Fister, won third place in the Hewlett Foundation's Essay Scoring Technology Competition (Getting Smart Staff, 2012). EdX also announced that their product would be available outside their MOOC environment. Although EdX's application is not yet available for testing, three long-standing commercial Automated Essay Scoring (AES) applications have been tested and are established in the academic literature (Shermis, Burstein, Higgins, & Zechner, 2010).

Alternatively, Daphne Koller and Andrew Ng who are the founders of Coursera, a Stanford MOOC startup, have decided to use peer evaluation to assess writing. Koller and Ng (2012) specifically used the term "calibrated peer review" to refer to a method of peer review distinct from an application developed by UCLA with National Science Foundation funding called Calibrated Peer Review™ (CPR). For Koller and Ng, "calibrated peer review" is a specific form of peer review in which students are trained on a particular scoring rubric for an assignment using practice essays before they begin the peer review process. As a complicating matter, Koller and Ng cited Chapman (2001) which is one of the first studies to establish the literature on UCLA's now commercial CPR application. Although differences may exist between Coursera's implementation of calibrated peer review and UCLA's product, UCLA's CPR has an established literature that allows it to be compared with AES applications. Thus, the purpose of this essay is to describe both AES methods and UCLA's Calibrated Peer Review™ program, provide enough pedagogical and mechanical detail to show the capabilities of these two methods of scoring and giving feedback on essays, and to provide a table and framework for comparing these two forms of assessing student writing in the context of MOOCs.

The Massive Open Online Course Environment

In this volume, Sandeen (2013) details the features, history, status, and challenges for MOOCs. When considering AES and CPR in MOOCs, several features of MOOCs are relevant. They:

- are web-based;
- are open enrollment and no-cost courses without enrollment caps;
- contain all the content or reference the freely available content required for the course; and,
- have very low instructor involvement from a student perspective after the course begins.

Many of the larger MOOCs have an enrollment that spans the globe (DiSalvo, 2012) and is educationally diverse (Educause, 2012). As Sandeen noted, only about 10% of the people who enroll in the largest MOOCs actually complete the course. MOOCs tend to follow the tenets of open education and provide frequent interactive activities during content presentation; are based on mastery learning which, among other things, provides practice and the ability to redo activities until the student is satisfied with their performance; and give feedback about attempts at activities. MOOCs run on a schedule with due dates, tests, activities, and other elements found in instructor-led online courses.

AES and CPR are different tools that can be used to assess writing in a highly automated course and have implications for the types of papers that can be scored, the consistency of feedback to students, the types of comments students receive, the need for instructor intervention and the range of what a student may learn in the course.

All of the features above will vary as more instructors develop new MOOCs, especially those instructors developing MOOCs outside the large consortia and on platforms with fewer controls such as Class2Go (an open source MOOC hosting product from Stanford). However, the features above describe the current state of MOOCs and are relevant for thinking about the ways AES and UCLA's CPR can be used in MOOCs. AES and CPR are different tools that can be used to assess writing in a highly automated course and have implications for the types of papers that can be scored, the consistency of feedback to students, the types of comments students receive, the need for instructor intervention, and the range of what a student may learn in the course. Sandeen (2013) shows that MOOCs have spurred experimentation with instruction; specific to this article, AES and CPR may become more accepted in the courses throughout the education continuum.

Automated Essay Scoring

On April 5, 2013, *The New York Times* website announced that EdX introduced an AES application that it will integrate within its MOOCs. Instructors reportedly will have to score 100 essays so that the machine learning algorithms can learn to score and give feedback on essays addressing a particular writing assignment. This type of technology for assessing students' writing is not new; the first successful AES system was programmed in 1973 but required punch cards and a mainframe computer, making it inaccessible to most instructors

(Shermis et al., 2010). As evidenced by MIT and Harvard's EdX announcement, this technology can now be applied to free online courses with enrollments over 150,000 students.

How Does AES Work?

Machine evaluation of essays correlated more highly with human raters of those essays than the human raters correlated with other human raters.

A more detailed treatment of AES mechanisms can be found in Shermis et al. (2010). To summarize, most AES applications build statistical models to predict human-assigned scores using features of essays that have been determined empirically or statistically to correlate with the ways humans rate those essays. Most AES models are built individually for each writing assignment or for a particular grade level. For example, the log of the number of words in an essay, when compared to other essays for that particular assignment or grade level, is one predictor of the score a human will assign to that essay. As an essay gets longer up to a point relative to the average essay length for that assignment or grade level, humans tend to score the essay higher. This simple example of a measureable characteristic of writing that predicts a human score is very rough. AES applications use many more machine-measured characteristics to more accurately predict human ratings of essays such as average word length, number of words in the essay, discourse element length, proportion of grammar errors, scores assigned to essays with similar vocabulary, and frequency of least common words. Moreover, some of these computed features are linked to particular feedback a human would give on an essay, so it is common for AES applications to score relative creativity, organization, and style and thus give feedback on these features of a particular essay as well as grammar and mechanics. Some AES applications use topical dictionary lookups for content specific to a writing assignment. Even more sophisticated Natural Language Processing computational elements are accessible to some AES applications such as text summarization, sentiment analysis, and semantic analysis. Three commercial systems currently dominate the AES market: e-rater™ made by Educational Testing Service (ETS) which is part of their Criterion℠ product,[1] Intellimetric™ made by Vantage Learning,[2] and Intelligent Essay Assessor™ made by Pearson Knowledge Technologies[3] (Graesser & McNamara, 2012; Shermis et al., 2010). Each of the three commercial AES applications uses some combination of the methods above. E-rater uses multiple linear regressions on at least 12 essay features to predict human scores by assignment or grade level. Intellimetric builds multiple statistical models from features of essays and pits the models against each other to get the best prediction of human scores. Intelligent Essay Assessor uses extensive topical dictionaries and different content feature measurements by topic to best predict human rater scores.

Does AES Work?

AES reached commercial viability in the 1990's by being indistinguishable from human evaluators for short essays with a specific focus (Attali, 2007). In a review of AES applications, Shermis et al. (2010) found that machine evaluation of essays correlated more highly with human raters of those essays than the human raters correlated with other human raters. That is, machine evaluation is distinguishable from human evaluation because it is more consistent than human evaluation. Moreover, AES detects differences in meaningful features of essays. Although each commercial product above uses different factors to rate essays, AES can detect and report about grammatical errors, word usage errors, sentence variety, style, text complexity, vocabulary, content alignment with existing texts, thesis statements, supporting ideas, conclusions, and irrelevant segments (Graesser & McNamara, 2012; Shermis et al., 2010). AES is not yet able to assess complex novel metaphors, humor, or provincial slang (Graesser & McNamara, 2012). However, AES offers immediate, consistent feedback to students about important elements of their writing.

The Limitations of AES

AES applications do not understand texts in the way humans do. As writing becomes more unique--such as in term papers on individually selected topics, academic articles, scripts, or poetry--commercial applications break down and currently cannot predict human

[1] http://www.ets.org/criterion
[2] http://www.vantagelearning.com/products/intellimetric
[3] http://kt.pearsonassessments.com/

scores (Graesser & McNamara, 2012). The National Council of Teachers of English (NCTE) has issued a position statement against machine scoring of student essays with an annotated bibliography. The reasons NCTE cited include the restricted range of essays AES is used on, vagueness of most AES feedback, and the potential that students and teachers who know AES will be used may turn writing for a machine into a game of correcting surface features and getting the correct length of essay rather than participating in a writing and learning exercise (National Council of Teachers of English, 2013). Further, although some of the recent literature on AES is very positive, it is dominated by results from industry (Crusan, 2010) which may not generalize to higher education. From an instructor's perspective, AES solutions all require training on human rated texts and often benefit from texts rated by multiple human raters and texts of significantly varying quality (Attali, 2007; Shermis et al., 2010). Even with EdX's announcement that an instructor will only need to grade 100 papers to train their application, 100 papers is a significant time investment. Lastly, a few studies suggest that structured, computer-regulated peer evaluation in specific situations may be more beneficial to students than just feedback on their writing (Heise, Palmer-Judson, & Su, 2002; Likkel, 2012).

Although some of the recent literature on AES is very positive, it is dominated by results from industry which may not generalize to higher education.

Calibrated Peer Review™, Version 5

UCLA's CPR is a stand-alone, web-based application that both manages the workflow for their specific peer review process and scores how well peer reviewers perform (see http://cpr.molsci.ucla.edu). CPR allows large numbers of students to:

- turn in essays,
- learn what the instructor believes are the critical points in those essays by scoring instructor-provided essays with a multiple choice rubric,
- perform peer review of their fellow students' work,
- perform a self-evaluation of their own work, and
- receive all the feedback from their peers who reviewed their work.

How Does UCLA's CPR Work?

Students complete four tasks which are scored when using version five of CPR. First, students write an essay which is scored by taking the weighted average of ratings given by three peer reviewers. Second, the students calibrate to the instructor's expectations by rating three essays provided by the instructor on a multiple-choice rubric. The instructor assigns a correct answer to each item on the rubric for each calibration essay and the students are scored by how well they match their instructor's answers. At the end of this task, students are assigned a Reviewer Competency Index (RCI) which functions as a weighting multiplier on the scores they have assigned to other students. Very low RCIs result in a 0 weight. Third, each student reviews three of their peers' essays with the rubric. The peer review task is scored by how well the individual reviewer's rating of the essay matches the weighted rating of the essay. Finally, students complete a self-evaluation of their own essay which is scored by how well they match their peers' weighted review scores.

Students receive feedback in CPR twice. First, during the calibration task, students get instructor-written feedback about the answers they chose on the multiple choice rubric for each training essay. The student may learn that they chose the correct answer for the rubric item or they may learn why the answer they chose was incorrect. Students who have not met the expectations for the calibration essay set by the instructor must retake that particular calibration trial a second time. In those cases, feedback is given again and the score on the second try stands. Second, students receive their peers' feedback on their essay from the peer review process including information from each rubric item weighted by the peer reviewers' RCIs. Thus, if three peer reviewers give differing answers on an item on the rubric (such as "were there more than three grammatical errors in the essay?"), the student with the highest RCI will be treated as providing the correct feedback and the other two as incorrect.

CPR also predicts potential scoring problems for the instructor. At the end of the assignment, the instructor gets a list of the essays that had three low RCI reviewers or had

fewer than three peer reviewers because of students dropping out of the assignment. Finally, in CPR version five, all the ratings and work the students do can be downloaded and mined with external tools.

Does CPR Work?

Since Russell, Chapman, and Wegner (1998), a mostly positive literature has been building about CPR. Studies that have examined student learning using CPR have found that CPR does result in learning the material students write about (Margerum, Gulsrud, Manlapez, Rebong, & Love, 2007; Pelaez, 2001; Russell, 2005), improves particular writing skills (Gunersel, Simpson, Aufderheide, & Wang, 2008), and improves related skills like the ability to evaluate material (Gunersel et al., 2008, Margerum et al., 2007; Russell, 2005).

Notably, there are few studies that compare traditional feedback on writing assignments with CPR. Hartberg, Gunersel, Simpson, and Ballester (2008) compared students' ability to write abstracts when students received TA feedback in a 2004 class and CPR in a 2005 class. Surprisingly, they found better student performance with CPR, speculating that there was a clearer connection between the instructor and student when CPR rather than TAs were used. Heise et al. (2002) found that students who received feedback by the instructor in a traditional way did not improve their writing and critical reasoning from assignment to assignment, but students who responded to an identical writing prompt and worked though the CPR process did. Likkel (2012) found that students who turned essays in to their instructor and received feedback did not gain a sense of confidence in evaluating their own writing, but students who followed the CPR process for the same assignment did. There are dissenting studies, however. Walvoord, Hoefnagels, Gaffin, Chumchal, and Long (2008) found that, although the scores students assigned to their peers' writing with CPR were comparable to those assigned by the instructor, there was no reported increase in student learning of content. Furman and Robinson (2003) reported no improvement on student essay scores in CPR throughout a course even though students perceived CPR as mostly helpful.

Students who received feedback by the instructor in a traditional way did not improve their writing and critical reasoning from assignment to assignment, but students who responded to an identical writing prompt and worked though the CPR process did.

Furman and Robinson (2003) also documented significant student resistance to using CPR. As a counterpoint, Keeney-Kennicutt, Guernsel, and Simpson (2008) described an instructor's continuous refining of her process to overcome student resistance to CPR using students' feedback and work with a faculty development team over eight semesters. Students were initially opposed to CPR, but Keeney-Kennicutt et al. (2008) showed significant gains in student satisfaction (shifting to overall positive attitudes) and students' belief that CPR helps them learn, write, and evaluate better. In this case study, the instructor:

- wrote her own assignments tailored to her course content and knowledge of her students rather than using the CPR assignment library;
- developed a detailed, 4-page handout (currently a 3-page handout attached to her syllabus is available at http://www.chem.tamu.edu/class/fyp/wkk-chem.html);
- framed her presentation of CPR to her students as an alternative to multiple choice tests for demonstrating their knowledge;
- offered to review the peer ratings for any student who asked; and,
- added an in class discussion of strategies for success on the calibration portion of the assignments.

These interventions significantly improved her students' experiences and resulted in students reporting that CPR is a useful tool. There are no published, comparable studies of student learning gains with continually refined CPR assignments over multiple semesters.

Limitations of CPR in a MOOC Environment

There are technical challenges for online courses with 150,000 students enrolled in them. Specifically for CPR, the basic system requirements (University of California, 2012) may not be sufficient for the load a large MOOC may generate. Thus, a professional technical

review and test of the server system housing CPR for a MOOC should precede enrolling students in such a course.

CPR's process may be difficult to scale to 100,000 students because some essays are scored only by three low RCI reviewers. This problem is dependent on student performance in the calibration phase and thus the number of essays with problems increases linearly with class size (assuming the quality of calibration performance stays constant for students as enrollment increases). Thus, if a MOOC has 100,000 students in it, and 10% finish the course (10,000), a 10% problem rate in CPR would translate to 1,000 essays with potential scoring problems. A potential solution to this problem would be to implement some training as Keeney-Kennicutt et al. (2008) reported. It may be possible to use mastery-based practice drills simulating the calibration phase outside of CPR so that students could attempt the drill over and over to master the process before they get into CPR. Most MOOCs do not reach the 100,000 student level; a 2,000 student MOOC may only have 20 problem essays.

Balfour (2011) noted three more relevant limitations of CPR. First, CPR relies on a web-based text input that requires basic HTML skill to format well. Many instructors provide an HTML tutorial either in writing or as a video.[4] Second, because CPR has a fixed rubric for each assignment, it is difficult to design a rubric for a term paper that allows students to use multiple outside references. Tightly focused essays with common sources fit better within CPR's model. Finally, there is a practical word limit when using CPR. Because students use the instructor's rubric on seven essays (three calibration essays, three of their peers' essays, and then once for a self-evaluation), essays containing more than 750 words can become difficult for students to manage. This limitation will depend more on the course expectations and level of students than the others.

This combination of AES and CPR may be very powerful and could produce stronger writers more efficiently than just human evaluation.

Comparing AES to CPR

Both AES and CPR have advantages and disadvantages in the context of MOOCs. Figure 1 offers a comparison of generalized AES methods of assessment and CPR.

Factor	AES	CPR
Types of Papers Scored	-Leveled or topical essays -Focused essays -Structured is better -More literal than figurative	-Single topic from common sources -Short essays -May be a little less structured -May be used for some figurative texts
Consistency of Scoring	-Highly consistent	-3 student raters provide feedback with visible disparities to the writer -Quality of calibrations and rubric partially determine consistency of score
Comments Provided	-Major element such as creativity, style, and organization -Based on statistical analysis or lookup -Likely to miss subtle elements	-May be enabled on every rubric element -Messy, human-based comments -Vary by reviewer ability and helpfulness
Instructor/TA Intervention	-Requires training essays: 100+	-CPR problem list may not scale up to multiple tens of thousands of students -Students often doubt peer assessment
Advantages for Student Learning	-Rapid feedback -Categorical and overall review	-7 uses of instructor rubric on content -Teaches evaluation skills -Self-evaluation after peer review -Required repetition/time on task

Figure 1. *A comparison of AES and CPR in MOOCs.*

Several types of written assignments are not likely to be successfully scored by either AES or CPR. The more unique or creative a piece is, the less likely that either method will produce a good evaluation. Although CPR can be used with more figurative and creative pieces, the length of the work is still a factor. Anecdotally, one instructor has successfully used computer assisted peer evaluation in a creative writing class, but the class was small and as intensive as similar creative writing classes (J. G. Smith, personal communication, 2010).

[4] See Jiffin Paulose's videos on YouTube.com; he is a biology instructor at the University of Kentucky

Conclusion

Both EdX and Coursera have announced ways that software will assist with written assignment in MOOCs, with EdX using an AES application and Coursera using a form of "calibrated peer review." While neither EdX's nor Coursera's tools have an established literature, both AES in general and UCLA's CPR specifically do. Automatic essay scoring has been commercially viable for more than a decade and can give nearly immediate feedback that reliably matches human raters for several types of essays. It can also give categorical feedback to students to help them improve their writing. UCLA's CPR makes writing possible in large section classes, gives human-generated feedback, and helps to train students in evaluation skills.

The AES literature is dominated by publications relying on corporate labs and data; this is in no small part because of the accessibility of large numbers of essays from commercial testing companies. MOOCs offer a new set of data for AES testing which has the possibility to substantially refine or change the state of that literature.

Instructors may favor one method or another when considering the way a MOOC should be structured. These decisions may be based on several factors such as the pedagogical outcomes of the particular method, the type of writing necessary in the MOOC, decisions about student tolerance for peer commentary on their work, and the work load the method might produce. However, in the spirit of experimentation in MOOCs noted by Sandeen (2013), a writing-based MOOC might use AES for giving students feedback on multiple rounds of drafts, but then use CPR for final evaluation. With this model, it is possible that the more mechanical writing problems could be corrected earlier in the writing process, improving the quality of essays feeding into CPR. Subsequently, students using CPR to review essays may be exposed to higher quality writing and thinking which may, in turn, benefit them even more than using CPR with lower quality essays. This combination of AES and CPR may be very powerful and could produce stronger writers more efficiently than just human evaluation.

As previously noted, Crusan (2010) stated that the AES literature is dominated by publications relying on corporate labs and data; this is in no small part because of the accessibility of large numbers of essays from commercial testing companies. MOOCs offer a new set of data for AES testing which has the possibility to substantially refine or change the state of that literature.

Finally, with the current technology, some types of writing are probably outside the reach of MOOCs. There is no literature to suggest that either AES or CPR can accurately assess figurative or creative pieces, or original research pieces. Some type of peer review software that relies heavily on the students being closer to experts in their own right might bring these types of writing into larger courses; but, not every undergraduate course that uses writing as a form of assessment will translate to the MOOC format.

References

Attali, Y. (2007). *On-the-fly customization of automated essay scoring* (RR-07-42). Princeton, NJ: ETS Research & Development. Retrieved from http://www.ets.org/Media/Research/pdf/RR-07-42.pdf

Balfour, S. P. (2011). Teaching writing and assessment skills: The intrinsic and conditional pedagogy of Calibrated Peer Review™. In Flateby, T. (Ed.), *Improving writing and thinking through assessment* (pp. 211-223). Charlotte, NC: Information Age Publishing.

Crusan, D. (2010). Review of Machine scoring of student essays: Truth and consequences. *Language Testing, 27*(3), 437-440.

DiSalvio, P. (2012). Pardon the disruption: Innovation changes how we think about higher education. *New England Journal of Higher Education.*

Educause. (2012). What campus leaders need to know about MOOCs. Retrieved from http://net.educause.edu/ir/library/pdf/PUB4005.pdf

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.

Getting Smart Staff. (2012, May 9). Hewlett Foundation announces winners of essay scoring technology competition Retrieved from http://gettingsmart.com/2012/05/hewlett-foundation-announces-winners-of-essay-scoring-technology-competition/

Graesser, A. C., & McNamara, D. S. (2012). Automated analysis of essays and open-ended verbal responses. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (pp. 307-325). Washington, DC: American Psychological Association.

Gunersel, A. B., Simpson, N. J., Aufderheide, K. J., & Wang, L. (2008). Effectiveness of Calibrated Peer Review™ for improving writing and critical thinking skills in biology undergraduate students. *Journal of the Scholar ship of Teaching and Learning, 8*(2), 25-37.

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.

Heise, E.A., Palmer-Julson, A., & Su, T.M. (2002). Calibrated Peer Review writing assignments for introductory geology courses. *Abstracts with Programs (Geological Society of America), 34*(6), A-345.

Keeney-Kennicutt, W., Guernsel, A. B., & Simpson, N. (2008). Overcoming student resistance to a teaching innovation. *Journal for the Scholarship of Teaching and Learning, 2*(1), 1-26.

Koller, D., & Ng, A. (2012). *The online revolution: Education at scale* [PowerPoint slides]. Retrieved from https://www.aplu.org/document.doc?id=4055

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

Margerum, L. D., Gulsrud, M., Manlapez, R., Rebong, R., & Love, A. (2007). Application of Calibrated Peer Review (CPR) writing assignments to enhance experiments with an environmental chemistry focus. *Journal of Chemical Education, 82*(2), 292-295.

Markoff, J. (2013, April 4). Essay-grading software offers professors a break. *The New York Times.* Retrieved from http://www.nytimes.com/2013/04/05/science/new-test-for-computers-grading-essays-at-college-level.html

National Council of Teachers of English. (2013). Machine scoring fails the test. *NCTE Position Statement on Machine Scoring.* Retrieved from http://www.ncte.org/positions/statements/machine_scoring

Pelaez, N.J. (2001). Calibrated peer review in general education undergraduate human physiology. In P. A. Rubba, J. A. Rye, W. J. DiBiase, & B. A. Crawford (Eds.), *Proceedings of the Annual International Conference of the Association for the Education of Teachers in Science* (pp. 1518-1530). Costa Mesa, CA.

Russell, A. A. (2005). Calibrated Peer Review™: A writing and critical-thinking instructional tool. *In Invention and impact: Building excellence in undergraduate science, technology, engineering and mathematics (STEM) education.* Washington DC: American Association for the Advancement of Science.

Russell, A., Chapman, O., & Wegner, P. (1998). Molecular science: Network-deliverable curricula. J*ournal of Chemical Education, 75*, 578-579.

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

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* (3rd ed., pp. 75–80). Oxford, England: Elsevier.

University of California. (2012). Calibrated Peer Review: Web-based writing and peer review. Retrieved from http://cpr.molsci.ucla.edu/SystemRequirements.aspx

Walvoord, M. E., Hoefnagels, M. H., Gaffin, D. D., Chumchal, M. M., & Long, D. A. (2008). An analysis of Calibrated Peer Review (CPR) in a science lecture classroom. *Journal of College Science Teaching, 37*(4), 66-73.