Scaling Up Student Assessment: Issues and Solutions

Paul J. A. van Vliet
University of Nebraska at Omaha

Online courses permit the enrollment of large numbers of students, which forces instructors to address the problem of providing valid and reliable assessments of student performance on a large scale. This paper examines two broad approaches for scaling up student assessment and feedback in higher education: automated assessment techniques and distributed assessment methods.

INTRODUCTION

Online education offers educational institutions a wealth of advantages: by offering courses online, universities are able to attract students worldwide, while also better serving local students whose schedules prohibit attending regular class sessions. Online courses as a rule do not require the use of physical classrooms and hence can accommodate larger numbers of students, resulting in potential educational efficiencies. However, for the faculty teaching these online courses, larger enrollments present a substantial challenge: how to assess student achievement and provide effective feedback on student work as the number of these students grows substantially. This is particularly relevant in Massively Open Online Courses (MOOCs) which permit the enrollment of hundreds or thousands of students in a single course section. (Balfour, 2013; Solomon, 2013) Students enrolled in MOOC-bases courses increasingly wish to validate their learning or obtain some form of recognized academic credit, requiring these courses to include valid and reliable student assessments. (Sandeen, 2013) As established universities such as Harvard and MIT start to offer MOOC-type courses with large enrollments, their instructors are finding that they "have limited assessment capabilities and grading options compared to residential courses," as most areas of higher education have not yet established large-scale assessments. (Ho et al., 2014)

Just how can instructors scale up assessment and feedback efforts while maintaining high levels of quality and academic rigor? It is commonly feared that scaled-up approaches to student assessment will result in inferior evaluations of student work and lower quality feedback to the student (Kulkarni, 2014), but is this so?

This paper examines current efforts to scale up student assessment and feedback in higher education. In the following sections, the purpose of student assessment is reviewed, and criteria for quality assessment practices are identified. The paper will then evaluate the pros and cons of existing methods for large-scale assessment as well as consider some promising innovative approaches.

STUDENT ASSESSMENT AND EVALUATION – CRITERIA FOR QUALITY

It goes well beyond the objectives of this paper to summarize the vast field of knowledge related to student assessment. However, for the purposes of evaluating the various methods for scaling up student assessment that will be discussed in this paper, a shared understanding of basic terms and criteria needs to be established

Student assessment commonly refers to a "wide variety of methods that educators use to evaluate, measure, and document the academic readiness, learning progress, and skill acquisition." (Assessment, 2013) We generally distinguish among pre-assessments which measure student knowledge and skills before they begin a lesson, course, or academic program; formative assessments which measure the ongoing progress of students as they proceed through a lesson or course; and summative assessments which measure the student's performance at the end of a learning unit or course. (Assessment, 2013; Carlson, Berry & Voltmer, 2005; Lines, 2004) Within educational institutions, student assessment efforts serve three broad purposes: they support the student learning process, permit the formal certification of student achievements, and provide for monitoring and accountability of the educational process to outside stakeholders. (Harlen, 2007; Lines, 2004)

When developing the student learning process, the design of instruction and assessment are closely interrelated. The content that is presented through the various methods of instruction (such as lectures and readings) is reinforced when assessment methods (such as quizzes, exams, papers, and other assignments) are specifically designed for students to internalize this content. (Haber, 2013) In addition, educators have long known that assessment methods must be carefully designed, as what these methods measure has a great impact on what students attempt to learn. For example, if an assessment only measures a student's factual knowledge, these students will focus on learning facts. If the instructor aims to assess other student skills or competencies, the deployed assessment methods should reflect this. (Baartman et al., 2006)

Most assessments – particularly at the university level – often have substantial consequences for the students being assessed, such as when a failing grade requires a student to repeat a course. Consequently these assessments have to be both credible and trustworthy. A good assessment method must therefore have both validity (it measures accurately what it proposes to measure) and reliability (the measure is consistent or reproducible across time, measurements, and instructors). (Harlen, 2007; Jonsson & Svingby, 2007; Lines, 2004) Such qualities are most easily achieved for objective tests, such as math problems or the direct recall of facts, but should extend to the more subjective assessments of student papers and designs. (Lines, 2004)

While validity and reliability are the most common criteria for the evaluation of assessment methods, the evaluation of these methods could also include their efficiency (the time and resource requirements of the assessment method), fairness (to show the absence of bias toward certain groups of students), impact (the accuracy of the consequences of an assessment in relation to what was actually measured), meaningfulness (the perceived value of the assessment task to the student's goals and interests), and transparency (the clarity of the assessment and its scoring criteria to the student). (Baartman et al., 2006; Harlen, 2007)

In the end, the design of the assessments and the quality of the feedback these assessments provide to the students will greatly influence just what students may learn in a course. (Balfour, 2013) Consequently these assessments ought to be carefully designed. Any tactic for scaling up these assessments should take into account the vital role that assessment plays in the learning process. Two broad approaches for scaling up student assessment will be discussed in the subsequent sections. The first approach aims to automate the scoring and feedback process; the second aims to involve students as peer graders to make the scoring and feedback process manageable.

AUTOMATED ASSESSMENT TECHNIQUES

An obvious idea for scaling up student assessment is through the use of information technology. This section will first describe traditional approaches for automating student efforts and will then detail the use of Automated Essay Scoring technologies.

Traditional Automated Assessment Techniques

The use of information technology to assess student learning is commonly referred to as Computer Assisted Assessment (CAA). This technology has long been deployed to score objective tests, such as multiple-choice exams, true-false exams, etc. Early CAA applications used optical scanning to read and process student answer sheets against an instructor's grading key. The use of direct student entry at a computer – now commonly using web-based testing applications – has expanded the range of question types possible in computer-scored exams. Computer Assisted Assessment offers some powerful advantages. Automated assessments can provide immediate and detailed feedback to the students about their learning progress and can help identify areas where students should improve their efforts. Resource-wise, automated assessments save instructors substantial time from having to grade exams manually. In addition, cumulative statistics can be quickly and easily compiled for aggregated evaluations of individual students or for a course as a whole. (Lines, 2004)

Because of these advantages, MOOC-type courses have deployed automated assessments from the start, particularly for objective testing. (Sandeen, 2013) While multiple-choice exams remain common, the range of assessments now includes formulaic problems with specific correct answers, the explication of logical proofs, and vocabulary and short-answer testing. In addition, automated assessments are now used for computer programming assignments to check the accuracy of student program algorithms or output. (Balfour, 2013)

The multiple choice exam, however, remains the most widely used type of objective testing. The speed and efficiency of exam administration, scoring, and analysis are the best known advantages of the multiple choice exam (Lines, 2004; Morrison & Free, 2001), but additional advantages exist. Multiple choice exams offer absolute reliability when administered and graded by computer. This type of exam can pose a large number of questions in a short amount of time, providing for broad subject coverage per exam. Multiple choice questions can be drawn from a pool, can be reused as often as needed, and can be presented in random order so as to reduce student cheating. (Lines, 2004)

The use of online learning environments permits instructors to further improve the use of multiple choice exams, and to enrich the learning experience. For example, online lessons are now often augmented with automated multiple-choice quizzes to assess a student's retrieval learning. This practice enhances the long-term memory of a lesson's facts by recalling this information from short-term memory. This approach is based on the idea that the act of "retrieval" reinforces the content in the learner's memory. Studies have found this to be an effective practice for enhancing the long-term recall of information. (Glance, Forsey, & Riley, 2013) The online lessons, and even the quizzes themselves, can be enriched with multimedia content, such as the inclusion of short videos. However, Lines (2004) cautions that instructors should be cautious as the assessment may lose validity when it tests a student's computer or software skills more than the lesson's original subject.

Some online courses permit students to repeat assignments until passing or perfect scores are achieved; something which can only be possible using automated grading. The immediate feedback provided permits students to continue to work on a problem until they solve it correctly or understand its subject matter well enough. The drawback to this approach is a lack of differentiation in summative assessment of student performance, as it would permit all sufficiently motivated students to obtain full or passing credit. (Solomon, 2013)

Multiple choice exams and other common automated grading systems are not without their drawbacks, as critics have noted a lack of substantial feedback and a limited view of learning.

While pre-written feedback is now commonly included in automated testing systems, such feedback is not individualized to the student and is therefore incapable of responding to a student's argument for

selecting a specific answer on a multiple choice exam. (Solomon, 2013) This drawback is not limited to these exams; the aforementioned assessments of computer programming assignments are capable of testing algorithms for correct output but would not be able to offer substantive feedback about the efficiency or structure of a student's code. As Solomon (2013) noted: "everybody knows that wrong answers taken many shapes and sizes," and providing custom feedback remains outside the capabilities of automated assessment for now.

Computer Assisted Assessment systems have also been critiqued for focusing on a shallow approach to learning. In particular, the focus of multiple choice exams on information transfer and student recall of facts (the aforementioned "retrieval" practice) is viewed by some educators as being insufficient for a substantial assessment. Nielson (2014) notes that students can best internalize information through the process of applying it, which isn't enabled by most multiple choice exams. Lines (2004) too notes that it is difficult to test higher order skills using automated assessments as students aren't permitted to demonstrate their communication or original thinking abilities.

While substantial feedback remains an important hurdle, the issue of higher order learning and critical thinking has been addressed specifically by various researchers and educators. They argue that while it requires substantial effort, it is possible to write substantive multiple choice questions which test higher-order thinking and concept application. (Lines, 2004; Morrison & Free, 2001)

Carefully constructed multiple choice questions could require the direct application of course content, the interpretation of information, the analysis of an explicit problem, or the evaluation of a set of more complex alternatives. Particularly when the students are required to distinguish among plausible alternative alternatives for a question or are asked to identify the best, most important, or highest priority option, critical thinking is encouraged. Well-written questions could also require the students to apply multilogical thinking, which is defined as "thinking that requires knowledge of more than one fact to logically and systematically apply concepts to a [...] problem" (Morrison and Free, 2001) For example, in a high-enrollment online course on the subject of the history of popular music, automated exams include questions which first require students to listen to short pieces of music. These pieces are then associated with multiple choice questions, requiring the students to thoughtfully consider their answer while still providing the instructor with an automated grading option. (Roland, 2015) Given that especially MOOC-type courses rely to a great extent on multiple choice exams for assessment, Haber (2013) insists that educators should spend the additional effort required to ensure that these assessments are both challenging and valid.

MOOC providers, however, have not limited their automated assessment efforts to these traditional types. In recent years these organizations have experimented with and prototyped a variety of technology-based assessments to scale up assessment efforts. (Sandeen, 2013) Much of this effort has targeted the use of computer algorithms to assess student writing.

Automated Essay Scoring (AES)

It is not always possible to properly assess student knowledge, skills, or other aspects of learning using the aforementioned traditional automated assessment techniques. Particularly in higher education, course grades often result from subjective evaluations or expert assessments by course instructors, especially for writing and design assignments. For such work, there is no clearly defined "right answer" that can easily be determined automatically. This presents a substantial challenge to the scaling-up of student assessment, and could inhibit the development of MOOC-type courses in certain disciplines. (Duhring, 2013)

As course enrollments become very large, instructors simply can't review student essays or other open-ended work as they do in smaller courses. As Watters puts it, "grading essays is incredibly time-consuming" and "giving meaningful feedback on student writing (and by extension on student thinking) is hard work." The process of grading essays requires the grader to respond to content, form, and mechanics of the writing, particularly when the goal of such individualized feedback is to "help a student learn, help the student be a better writer and a better thinker." (Watters, 2012)

This assessment challenge has been answered with the development of Automated Essay Scoring (AES) applications. These systems have the potential to reduce the cost and time which are otherwise associated with the human effort of reading, interpreting, and rating student writing. (NCTE, 2013)

The AES process begins with the development of a "training set" of instructor-scored essays. For example, an Automated Essay Scoring application designed for MOOC courses offered by EdX first requires instructors to manually score 100 essays so that its machine learning algorithms can learn to score and give feedback on specific essay assignments. Using these scoring examples, AES applications can learn to evaluate essay length, grammar errors, average word length, specific vocabulary usage, word frequency, etc. In essence, AES algorithms extract those features that instructors have rewarded with high scores from the "training set" and subsequently build a statistical model which predicts human-assigned scores using those features. More advanced systems deploy Natural Language Processing techniques, such as text summarization, sentiment analysis, and semantic analysis. (Balfour, 2013; Kolowich, 2014)

It is important to realize that AES applications don't actually "read" student essays the way human graders do, and the applications do not comprehend the actual meaning of words, sentences, and essays. Instead these applications "describe" the student essays, compiling lists of every feature, word, phrase, or syntactic structure in each essay, and then comparing these findings against the features extracted from the essays scored by human instructors. (Mayfield, 2013; Reich, 2012)

While AES applications were originally developed in the 1960s, they did not achieve broad usage or commercial viability until the 1990s. (Attali, 2007) The more recent creation of MOOC-type courses has spurred substantial development in the AES field. The major providers of these courses – Coursera, EdX, and Udacity – are all making considerable investments to further develop the features and sophistication of AES applications. In addition, the Hewlett Foundation encouraged innovation in this field by reaching out to data scientists and machine learning specialists through its sponsorship of the Automated Student Assessment Prize. Its first phase, which focused on Automated Essay Scoring, attracted approximately 150 contestants. (Markoff, 2013; Vander Ark, 2012)

The Benefits of Automated Essay Scoring

The main advantage of AES applications is of course grading speed. Human graders – even when spending only a few minutes evaluating a short essay - may grade approximately 30 writing samples per hour. EAS applications can grade 16,000 such essays in 20 seconds. (Winerip, 2012) Moreover, for short essays with a very specific focus, AES ratings have been found to closely match human grading efforts and to apply such assessments objectively and consistently, at times even more so than human graders would. (Balfour, 2013; Reich, 2012; Shermis & Hammer, 2012)

The use of AES applications provides additional potentially significant educational benefits. As these applications are refined and improved, course instructors and graduate assistants will be able to spend far less time grading assignments and more time interacting with students. The time saved by AES applications could permit course staff to better tutor or mentor students in order to advance student learning outcomes. (Boston & Stephens-Helm, 2012)

A second educational benefit is that the use of AES applications would permit instructors to assign more writing assignments per course than would be possible if they had to grade these manually. This means that students get to practice their writing skills and receive feedback on their writing more frequently. Students may even be able to submit drafts to an AES application in order to obtain feedback while they are still working on the writing assignment: an automated instance of formative assessment. Through this repeated and sustained practice, similar to practicing math drills, students have the opportunity become better writers and thinkers. Even if AES feedback focuses mostly on grammar, vocabulary, and sentence structure, many students would benefit from improve their writing skills in these areas. The instructor could then focus on mentoring students on style, logic, and content. (Glance, Forsey, & Riley, 2013; Mayfield, 2013; Vander Ark, 2012)

Limitations and Drawbacks of Automated Essay Scoring

While AES applications have to potential to solve a large problem in the scaling-up of student assessment, at this time they are subject to some inherent limitations and have received some strong criticisms.

A significant limitation of AES applications is that currently they are most effective at grading short, focused, and predictable student essays. This presents a problem for educators in higher education who need to grade student writing that includes innovative ideas, advanced research, creative or poetic expression, etc. It should be understood that AES applications simply do not "understand" a text in the same way that human readers do. The AES algorithms are incapable of recognizing complex arguments, complex metaphors, innovative content, the inclusion of humor, or a highly individual writing style, and as such are limited in their applicability at the university level. (Balfour, 2013) It should be noted, though, that AES developers generally do not claim AES suitability for such writing. (Mayfield, 2013)

Compared to human readers, many AES algorithms consider the assessment of student writing in rather simplified terms. While comparative studies have found that AES scores closely match human-generated essay scores (Shermis & Hammer, 2012), these findings have been criticized for being limited mostly to grammatical aspects of the student essays, rather than on writing skills such as organization, argument, and meaning. (Perelman, 2013) The essays needn't even be factually true, as EAS applications can't identify the veracity of the sentences they process. Errors of fact are overlooked in favor of well-structured sentences. (Winerip, 2012)

Once students receive AES-produced feedback and become familiar with the specific scoring criteria deployed by an AES, they might actually be able to "game" the software in order to obtain higher scores. For example, AES applications have been found to attach higher ratings to longer sentences, paragraphs, and essays, as well as to the use of more complex words. (Winerip, 2012) Students would be able to intentionally write toward these criteria. Les Perelman, who has taught writing at the Massachusetts Institute of Technology and is an outspoken critic of AES usage, has taken this idea a step further with the development of a tool called the Basic Automatic B.S. Essay Language Generator (or Babel), which generates essays based on a few keywords. These essays contain grammatically correct sentences but no coherent meaning, yet are able to obtain high scores from AES algorithms. (Kolowich, 2014)

While most arguments regarding AES applications have focused on the nature and accuracy of the scores provided by these applications, the specific nature of the feedback provided by these applications has also been criticized. Human feedback on student work is can be more than a grade and some comments about grammar; it can be "substantive, smart, caring, thoughtful, [and] difficult." (Watters, 2012) For open-ended student work, such feedback can be quite valuable to the student. A study by Dikli (2010) looked at differences in the feedback students received after writing personal essays. This study found that AES feedback tended to be lengthier than human feedback, more often redundant or repetitive, and at times so vague or generic as to be usable for the student. Students also found it lacked human interaction and personalized positive reinforcement. Human feedback tended to be shorter as it gave cumulative assessment about the essays as a whole, rather than address each sentence issue separately. Human feedback was also more specific and more consistently applied, and as a result was considered more useful by its recipients. Students also recognized the value of personalized positive reinforcement that was included in human feedback. As one student stated "Computer can't answer my questions. I can't ask 'what do you mean?'" (Dikli, 2010)

In light of these limitations of AES-based grading, the National Council of Teachers of English (NCTE) has issued a position statement which strongly opposes the use of automated scoring of student essays. The NCTE argues that AES applications are unable to assess or even recognize key aspects of good writing, such as organization, accuracy, relevance, and writing style. Instead these automated scoring systems favor more "shallow" aspects of writing such as grammar, spelling, punctuation, or word-length. Knowledge of AES algorithms could result in these systems being "gamed" by students as they write specifically towards automated scoring evaluations, rather than participating in the actual writing and learning exercise. (Balfour, 2013; NCTE, 2013)

Finally, instructors must realize that the use of AES applications is not effort-free and that they are not applicable to all courses which involve writing assignments. Many AES algorithms rely on machine-learning techniques which require the detailed human grading of a large set of sample essays in order for the algorithms to extract relevant features. (Balfour, 2013) As Mayfield (2013) points out, investing the time it takes to grade one hundred (or more) sample essays makes no sense for courses with small enrollments, such as university seminars. Only when enrollments grow to many hundreds or thousands of students per course does the investment in AES grading make sense. (Mayfield, 2013)

Assessing Automated Essay Scoring

Like other disruptive technologies, the use of Automated Essay Scoring has both passionate supporters and vocal critics. Supporters point to the efficiencies offered by AES to very large courses, and the detailed feedback provided to the students, particularly when directly addressing basic writing skills. Critics highlight the poor assessment of student essay content, the possibility of students gaming AES systems, and the inability of these systems to recognize creativity and innovation. While critiques regarding creative and innovative content are correct, it must be noted AES developers never designed these algorithms for that purpose. (Mayfield, 2013) The core of this argument actually relates to a validity issue regarding AES: does AES intend to measure basic student writing skills such as grammar, vocabulary, and syntax or does it intend to measure the quality or creativity of the essay contents? AES provides reasonably valid measures of the former, and reliably so, but not of the latter. However, considering current investments in AES (such as the aforementioned Automated Student Assessment Prize), one would expect that AES algorithms and features will be enhanced over time. As natural language processing technologies progress, AES algorithms will likely be capable of tackling more substantial aspects of writing.

TABLE 1
EVALUATION OF AUTOMATED ESSAY SCORING

Assessment Criteria	Performance of Automated Essay Scoring
Validity	Reasonably good for assessing basic writing skills, but at this time still limited at assessing content and creativity.
Reliability	Algorithms ensure high levels of consistency across student essays.
Efficiency	For very large courses, efficiencies exist, but a substantial grading effort is required to train the AES algorithms.
Fairness	The emphasis of technical writing skills over essay content may unfairly target students for whom the writing assignment is not in their native language.
Impact	This is situation-dependent on the essays being graded by AES systems.
Meaningfulness	This is dependent on the nature of the assignment provided by the course instructor.
Transparency	AES systems may be overly transparent when students are capable of "gaming" the algorithms.

DISTRIBUTED ASSESSMENT METHODS

An alternative to the use of automation for scaling up student assessment is the practice of dividing up the work among multiple graders. This section will first describe why human grading continues to be needed and will then describe the use of the Calibrated Peer Review as a distributed assessment method.

The Continued Need for Human Reviewers

While automated assessments can be quite effective in some disciplines, others (such as design-oriented fields like architecture, product design, and software design) rely on the qualitative assessment of open-ended work. Moreover, the process of viewing and offering critiques of others' work is often a key aspect of the pedagogical approach of these disciplines. In these fields, successful educational efforts require that open-ended design work is assessed qualitatively as singular correct solutions often do not exist. Students in these fields need both qualitative formative feedback (that is, ongoing feedback as students are performing work) and summative feedback (the assessment of student learning at the end of an assignment or course). Given that automated grading systems usually do not capture the semantics of student work, such open-ended assignments usually rely on human graders. (Carlson, Berry & Voltmer, 2005; Kulkarni et al, 2013; Shah et al., 2013)

When such courses experience large enrollments, the human grading effort required is often prohibitive to their instructors. In some cases instructors have graduate assistants to help them with this effort. However, the availability of a sufficient number of graduate assistants to make the grading effort manageable cannot always be guaranteed. A potential solution to this problem is to have the students evaluate each other's work, an approach often referred to as peer grading. This practice could enable large courses to include student assignments that are impractical to grade automatically. Kulkarni et al. liken peer grading to crowdsourcing, which refers to the efforts of online communities to accomplish substantial tasks through the small, voluntary contributions of its members. Crowdsourcing participants tend to perform best when they are intrinsically motivated by the importance of the task at hand. In such cases, these crowd-workers tend to be receptive to short, well-designed training exercises, and their subsequent ratings can help provide a good indicator of quality. (Kulkarni et al, 2013)

The Calibrated Peer Review

The best known peer grading procedure is the Calibrated Peer Review, which requires participating students to be trained on a particular scoring rubric for an assignment, in order to ensure that student grading efforts are consistent with the grading practices and standards of the course staff. (Balfour, 2013; Carlson & Berry, 2003; Koller & Ng, 2012; Kulkarni et al, 2013)

The Calibrated Peer Review commences with the development of a specific scoring rubric for each course assignment. A rubric is generally understood to be a simple assessment tool that is designed to provide a qualitative rating of a student's work. A rubric does so by explicitly defining criteria for performance levels for each aspect or characteristic of the assignment. As such, a rubric makes clear - to both the instructor and the student - what aspects of the assignment are considered important and what to look for when assessing the work. By having both assessment criteria and performance levels explicitly stated, the use of rubrics aims to set valid and reliable standards for the assessment of student performance in a way that can be communicated effectively to the student graders. (Boston & Stephens-Helm, 2012; Jonsson & Svingby, 2007)

For each individual assignment that requires peer-grading, the students are first shown the scoring rubric, often with an explanation of its usage. Using this rubric, the students are then given a practice assignment to grade. Following this effort, the students are shown how course staff had assessed the same assignment so that the students can then calibrate their own grading practice. This calibration phase reduces concerns that assessment criteria become inconsistent or imprecise, and reliant on individual judgment rather than clear and consistent academic standards. (Freeman et al, 2012) Once this training phase has been completed, each student is usually asked to review the work of approximately five fellow students. (Duhring, 2013) In some instances, the process concludes with the students reviewing their own assignment using the scoring rubric as a means of self-assessment using the insights gained from the review process. (Carlson & Berry, 2003)

The use of peer grading changes the role of the instructor in a course. In traditional grading practices, the instructor's efforts are focused on doing the grading activity. When deploying peer assessment, the instructor not only has more time available to respond to student questions and mentor these students, but he or she also needs to spend more time presenting assessment criteria for the students to use. Such

presentations - usually in the form of rubrics - need to be constructed carefully and updated as the instructor gains experience with the course, the students, the student work, and the peer grading process. (Kulkarni et al, 2013)

The information technology applications already in place in many educational institutions can be used to execute the peer review process. Popular Learning Management Systems such as WebCT™ and BlackBoard™ make it convenient for instructors to gather student work, display rubrics, distribute assignments for grading, and collect peer rating documents. (Carlson, Berry & Voltmer, 2005) In addition, a study by Freeman et al. demonstrates a custom web-based application used to perform the Calibrated Peer Review process. (Freeman et al, 2012)

Benefits of the Calibrated Peer Review

The Calibrated Peer Review has received substantial attention from researchers. For example, a comprehensive study by Kulkarni et al. investigated peer grading using the Calibrated Peer Review process which included course staff feedback during the training process. Each student grader would subsequently grade five peer assignments. As a result, each student assignment would be graded by four or five randomly selected peer raters. As part of the research study, course staff would grade one out of five of each rater's assignments to assess the quality of the rater's grading. The study found substantial agreement among student and staff ratings. The study also found that time, familiarity and practice with peer assessment resulted in students producing higher quality assessments. (Kulkarni et al, 2013)

The peer review process – with its explicit rubrics and calibration phase – also results in the use of consistent assessment standards to student work. (Freeman et al, 2012) Such consistency is not only important among raters in a single course, but also among multiple course offerings and instructors, ensuring a consistent level of assessment quality over time and across course sections.

An additional benefit is the extensive amount of data that is collected by the peer review process regarding student work and assessment. This data can offer valuable insights in the measurement of learning outcomes in a course or curriculum. (Carlson & Berry, 2003)

The effective scaling-up of the grading process is not the only benefit of the Calibrated Peer Review process; researchers have also observed important pedagogical benefits for the students participating in peer reviews. The continued student engagement with a particular course assignment extends their learning experience as concepts are reinforced through the exposure to the rubric and to other students' work. (Boston & Stephens-Helm, 2012) Carlson & Berry note that the Calibrated Peer Review can function as a "cognitive apprenticeship model," in which students (as apprentices) receive valuable mentoring from their instructors and obtain additional insights from their peers during the grading process. (Carlson & Berry, 2003)

While participating in this process, the students see the course work from an assessor's perspective, applying the instructor's rating criteria to the work of their peers. (Kulkarni et al, 2013) At the same time, the students – reading and evaluating the work of their peers – cannot help but see how their fellow students write and think, all of which may help them see the strengths and weaknesses of their own work. (Balfour, 2013; Carlson & Berry, 2003; Watters, 2012) In other words, the process encourages students to take greater responsibility for their own learning, which would include developing the ability to objectively assess their educational achievements. These reviews, then, may result in a valuable self-assessment of the student's own efforts and provide them with ideas and inspiration for areas for improvement, helping them become "more resourceful, confident, and higher achievers" (Glance, Forsey, & Riley, 2013; Kulkarni et al, 2013)

Drawbacks of the Calibrated Peer Review

Practical experience with the Calibrated Peer Review has shown that this method does have some significant shortcomings, which mainly affect the quality of the review process and the quality of the feedback the student receives.

The review process assumes that all participating students are capable, motivated, and well-intentioned. However, some students were found to not appreciate the additional course workload

(evaluating the calibration-phase assignments and the actual peer assignments, in addition to performing their own assignment) that is required by the CPR method. (Balfour, 2013) Poorly motivated student graders may produce hastily written, low-quality assessment efforts, which benefit neither the course nor the students being evaluated. (Kulkarni et al, 2013; Watters, 2012)

Researchers have also found that students may simply be unprepared or unqualified to perform the review process and provide meaningful feedback to their peers. They are not formally trained on how to properly and consistently assess student work, they may not be native English speakers and thus have language issues to overcome, and they may lack the necessary content expertise to assess the work of fellow students. (Boston & Stephens-Helm, 2012; Kolowich, 2014; Shah et al., 2013) These limitations would likely affect the validity of peer reviews.

The latter problem is particularly pressing in open-enrollment MOOC-type courses. The students enrolled in these courses may not actually be actual peers in an academic sense, that is, the background knowledge and level of skill may vary substantially among students. Such asymmetrical skill levels – in which some students may be novices to a task or discipline while others are experts – can result in low-quality peer reviews and general student discomfort or defensiveness during the overall peer review process. (Carlson, Berry & Voltmer, 2005) This asymmetry could be addressed through carefully constructed rubrics which focus the student grader's attention to specific aspects of an assignment, or by the instructor providing a list of common errors for the peer graders to recognize, but neither would guarantee a consistent review process. (Kulkarni et al, 2013; Kulkarni, 2014)

It should be noted that the lack of peer rater expertise is less of an issue in traditional university courses where enrollment is controlled through course prerequisites; students in such courses are more likely to have similar levels of skill and knowledge. (Duhring, 2013)

The quality of the feedback a student receives is important to the student's individual learning as well as to the success of the course overall. When feedback is provided by poorly motivated or unskilled peer graders, the recipients of this feedback may perceive it as uninformative, inaccurate, useless, or unfair, thus reducing their satisfaction with the peer review process and possibly prompting student complaints. (Gibbs, 2012; Kulkarni et al, 2013; Watters, 2012) In a course with a small enrollment, deficiencies in peer grading and the resulting student complaints can easily be handled by an instructor. When enrollment grows very large, these complaints may overwhelm the instructor. (Shah et al., 2014)

This problem is exacerbated by the fact that peer reviews are usually anonymous; the students being reviewed do not know which fellow students evaluated their work, while peer reviewers know that the students being evaluated can't trace reviews back to them. While this practice protects student privacy and prevents attempts at collusion and grade inflation, it may undermine the effectiveness of the peer review process.

When students don't know who has assessed their work, they can't ask for clarifications about the review, nor can they gauge the quality of the review based on the known expertise of the reviewer. Particularly when reviews are negative, students may not receive the actual feedback they need for their learning to advance. (Boston & Stephens-Helm, 2012; Larson, 2014; Watters, 2012) In addition, anonymity was sometimes found to bring out "Internet trollishness," in which student raters left nasty or inappropriate comments for their peers without any repercussions. (Gibbs, 2012; Watters, 2012) Such incidents reduce the recipients' trust in peer grading and raise the question as to whether peer reviews can work reliably in a setting where anonymity precludes the creation of a student community.

Assessing the Calibrated Peer Review

In the end, the Calibrated Peer Review has great potential for providing large numbers of students with qualitative feedback on their work. The process has been found to be an effective approach for providing context-appropriate evaluations for open-ended course assignments. (Kulkarni et al, 2013) However, this process is not effort-free. The instructor will need to carefully devise course assignments and accompanying rubrics, and will need to spend time supporting students through the calibration and grading phases. In addition, the course itself will need to be designed to include the peer review process as part of its activities. Finally, students participating in peer reviews need to be properly motivated and

capable of providing substantive feedback. Instructors may need to consider building incentives into the assessment process to encourage students to give high quality feedback to their peers. Some research did find that sustained practice with this approach will over time ease its drawbacks as instructors and students both become more familiar and comfortable with it. (Kulkarni et al, 2013)

However, the assessment of open-ended and competency-based student work complicates the determination of its validity and reliability. For traditional, fact-based assessments, the determination of their validity and reliability is broadly used as a measure of their quality, and sufficient experience exists to measure both aspects well. When the concepts of validity and reliability are limited to the attainment of repeatable, objective, and standardized outcomes, open-ended assessments are often found wanting. When no "correct answer" exists for an assessment, validity and reliability needs to be established through more subjective judgments, such as through human experts. (Baartman et al., 2006)

TABLE 2
EVALUATION OF THE CALIBRATED PEER REVIEW

Assessment Criteria	Performance of Calibrated Peer Review
Validity	Validity is difficult to establish outright, but may be supported through careful construction of rubrics and calibration of student grading efforts.
Reliability	Proper training and motivation of student graders is required to establish reasonable levels of reliability.
Efficiency	Even though students grade peer essays, instructors are still responsible for developing rubrics, calibrating student grading, and handling exceptions and complaints. All these activities may require a substantial amount of time.
Fairness	Fairness of the assessments is greatly dependent on student grading efforts, but may be encouraged through proper training and motivation.
Impact	This is situation-dependent on the essays being graded by the students.
Meaningfulness	The ability to compare their own work to that of others, and the opportunity for self-reflection may enhance the meaningfulness of the process for participating students.
Transparency	The availability of a grading rubric enhances the transparency of the peer grading process; however, the anonymity of the process could reduce the transparency of specific grading efforts.

INNOVATIVE TECHNIQUES AND IDEAS

Automated Essay Scoring and the Calibrated Peer Review are the predominant approaches currently in use for scaling up student assessment efforts. It will likely take substantial time and effort for alternative practices to be developed and validated. Fortunately educators and researchers continue to develop concepts for improved and innovative assessment practices. This section discusses four such concepts, three of which are aimed at improving the Calibrated Peer Review and one which targets the efficiency of providing feedback.

An innovative take on the peer review process is offered by Shah et al., who recognize that students often have difficulty assigning an absolute grade to a student essay, given that they are not trained reviewers. For the Calibrated Peer Review process to work properly, it is necessary that the students grade each other's work reliably and consistently, which has not always been found to happen. However, these students may have an easier time distinguishing better work from poorer work. Hence Shah et al. propose a variation on the Calibrated Peer Review process through the use of ordinal or comparative peer grading. In their proposed ordinal peer grading approach, students are asked to compare student essays (preferably using a simple pair-wise comparison) and to indicate which of these they consider to be the "better" essay.

Written feedback then could include a contrast of the strengths and weaknesses of the respective essays being compared. The essays a student evaluates in this manner could include some that have been previously graded by course staff for calibration purposes. All of the peer review comparisons can then be aggregated using a software application which assigns a final evaluation to each student. While the concept of comparative grading sounds appealing, evidence of its effectiveness is not yet available. (Shah et al., 2013)

Shah et al. also propose an alternative method to improve the level of accuracy of the peer review process when deploying absolute rather than comparative grading. Given that as course enrollments increase, the number of poorly assessed student essays increase proportionally, they propose the addition of a dimensionality reduction step to the peer review process. In this step, software clusters student essays based on similarity of content, or on the presence of specific predetermined features. Highly similar content or feature-sets, they argue, ought to receive the same grade. If one essay in each cluster is properly graded (either by an instructor or by aggregating peer reviews), then the other essays in the cluster should receive the same grade. While the authors note that current software capabilities should be able to accomplish this cluster analysis, the method has not yet been operationalized (Shah et al., 2014)

To ensure that students involved in a Calibrated Peer Review actually have the knowledge and expertise needed to perform quality assessments, perhaps the process should involve past students instead of current students. And so building on the idea of crowdsourcing as a solution for the assessment of large courses, Kulkarni et al. introduce the concept of Community TAs: unpaid volunteer teaching assistants, recruited from among the high-performing students in a previous administration of a specific course. These Community TAs would grade assignments, answer student questions, and help the instructor in improving the assignments. The use of Community TAs could therefore provide opportunities for both peer mentoring and peer assessment, while allowing for greater control over the quality of the peer grading effort. The feasibility of this approach, especially for very large courses, has not yet been established. (Kulkarni et al, 2013)

Finally, an approach which focuses on the efficiency of an instructor's qualitative feedback process is discussed by Mandernach and Garrett. Their suggestion starts with the realization that much feedback on student writing tends to be repetitive (addressing such issues as grammar, style, content organization, argument development, referencing techniques, etc.). As a result, an instructor will over time provide the same piece of qualitative feedback over and over again. They argue that these pieces of feedback – whether they be individual sentences or full paragraphs – be collected in a "feedback bank." These pieces are then associated with short keyboard-codes. When grading student work, these pre-written pieces can be quickly recalled and inserted in a student's evaluation. As the instructor gains experience with the use of the keyboard-codes and as the "feedback bank" grows in content, providing high-quality feedback to student work can be accomplished with substantial speed. (Mandernach & Garrett, 2014) While this idea targets the efforts of individual instructors, one could easily imagine the development of a large feedback bank for a course which would provide support for a peer grading process, particularly when this pre-written feedback is coupled with a list of common assignment errors as suggested by Kulkarni. (2014)

CONCLUSION – EVALUATING SCALED-UP ASSESSMENT PRACTICES

This paper was initiated by the question of how student assessments could be scaled up for courses at the higher education level, given that online courses permit for large enrollments as physical classroom size is no longer an issue. The author also wondered just how courses with extremely large enrollments – those taught as MOOCs – could reliably assess student knowledge and skills. It was surprising to find out that at this time objective multiple choice questions are still the most common assessment method used, and that only two additional approaches have gained traction with educators: automated essay grading and peer grading.

While both approaches have some advantages, none are perfect. Kolowich (2014) quotes Piotr Mitros, chief scientist at edX, who states that all assessment approaches have drawbacks as "machines cannot provide in-depth qualitative feedback," "students are not qualified to assess each other on some

dimensions," and "instructors get tired and make mistakes when assessing large numbers of students." (Kolowich, 2014)

The experience of edX is particularly relevant here, as this endeavor at creating university-level MOOC-type courses was founded by the Massachusetts Institute of Technology and Harvard University. (Ho et al., 2014) Its participants have found that "some fields have well established large-scale assessments, but most areas of higher education do not." Consequently they have recognized that educators and researchers "need to invest more in high-quality, scalable assessments, as well as research designs, including pretesting and experiments, to understand what and how registrants are learning." (Ho et al., 2014)

Given the importance of student assessments to both students and educational institutions – what Harlen (2007) refers to as "impact validity – it is important to provide high quality assessments and feedback, even when course enrollments grow very large. This effort is further complicated by two issues that are outside the bounds of this paper but which deserve mentioning: authentication and plagiarism. Authentication refers to the process by which the educational institution ensures that the person completing the course assignments and exams is the same person who enrolled in the online course and obtains credit for it. Various student authentication and online exam proctoring technologies are currently deployed. (Sandeen, 2013) Plagiarism is the practice of students submitting work that is not their own. While plagiarism can occur in all types of courses, Young (2012) notes that in MOOC courses, where students often face no consequences, plagiarism is all too common.

MOOC courses, as it turns out, are treated differently by students, particularly when enrollment is truly open and not tied to a student's admission to a university degree program. Such courses have a potential for what Ho et al. (2014) refer to as "non-certified learning." They point out that while MOOC courses commonly measure student learning through statistics regarding the number of students who have completed all course requirements (and hence received MOOC credentials, certificates, etc.), these numbers do not tell the whole story. Even students who enroll in a MOOC and browse through the content but do not take part in assignments or other certification activities might still learn particular knowledge they were hoping to obtain. In addition, some students who obtain certification through a MOOC may already have existing expertise on the MOOC's topic and have merely engaged with the MOOC in order to obtain certification at a low cost. In other words, "noncertified registrants may have learned a great deal from a course, and certified registrants may have learned little." (Ho et al., 2014)

For those higher education courses that offer student certification, whether they be MOOC-type courses or not, the quality of student assessment and feedback remains a substantial challenge and innovative approaches are badly needed. Kulkarni (2014) argues that perhaps educators should not look to scale up existing approaches but ought to instead develop brand new assessment techniques that only work for large enrollments: "Could we transform scale into an opportunity? Could we go beyond being there, and design social computing technologies to enable education that is impossible at smaller scales?" (Kulkarni, 2014) (The "beyond being there" comment is a reference to a 1992 paper by Hollan and Stornetta, who argued that rather than try to imitate physical proximity - which provides a richness of information through face-to-face contact - telecommunications research should develop tools that go beyond such expectations. These would be telecommunications tools that people would prefer to use even if they have the option of face-to-face communication. Kulkarni argues for similar transformative innovations in student assessment.)

In the end, the rapid development of a variety of online learning technologies presents a complex challenge to educators. As Solomon (2013) asked: "What traditions from centuries of brick-and-mortar teaching should be transferred online, and what should we throw out? What worked well about old teaching models, and what can be improved?" But these flows of innovation and change need not be a one-way street. Methods, techniques, and technologies that were developed for MOOCs and online courses could just as easy migrate into traditional classroom-based courses. (Sandeen, 2013) As educators and researchers progress their efforts at assessment techniques, all of education stands to benefit.

REFERENCES

Assessment (2013). In S. Abbott (Ed.), The glossary of education reform. Retrieved from http://edglossary.org/assessment/

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

Baartman, L.K.J., Bastiaens, T.J., Kirschner, P.A., & Van der Vleuten, C.P.M. (2006) The Wheel of Competency Assessment: Presenting Quality Criteria for Competency Assessment Programs. *Studies in Educational Evaluation*, 32(2006), pp. 153-170.

Balfour, S.P. (2013) Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review. *Research & Practice in Assessment*, Volume 8. Retrieved from http://www.rpajournal.com/dev/wp-content/uploads/2013/05/SF4.pdf

Boston, W. & Stephens-Helm, J. (2012) Why Student Learning Outcomes Assessment Is Key to the Future of MOOCs. *National Institute for Learning Outcomes Assessment*. Retrieved from https://illinois.edu/blog/view/915/84723

Carlson, P.A. & Berry, F.C. (2003) Calibrated Peer Review and Assessing Learning Outcomes. In *Proceedings of the 33rd Annual Frontiers in Education Conference*, November 5-8, 2003, Boulder, CO. IEEE Digital Library.

Carlson, P.A., Berry, F.C. & Voltmer, D. (2005) Incorporating Student Peer-Review into an Introduction to Engineering Design Course. In *Proceedings of the 35th Annual Frontiers in Education Conference*, October 19-22, 2005, Indianapolis, IN.IEEE Digital Library.

Dikli, Semire. (2010). The nature of automated essay scoring feedback. *CALICO Journal*, 28(1), 99-134.

Duhring, J. (2013) Massive MOOC Grading Problem – Stanford HCI Group Tackles Peer Assessment. *MOOC News & Reviews*. Retrieved from http://moocnewsandreviews.com/massive-mooc-grading-problem-stanford-hci-group-tackles-peer-assessment/

Freeman, M., Willey, K., Hancock, P., Howieson, B., Watty, K., Abraham, A., O'Connell, B., & De Lange, P. (2012) Using technology to improve peer review and collaborative conversations to benchmark academic standards. In *Proceedings of the 33rd Annual Frontiers in Education Conference*, October 3-6, 2012, Seattle, WA. IEEE Digital Library.

Gibbs, L. (2012) Peer Feedback: The Good, the Bad and the Ugly. Retrieved from http://courserafantasy.blogspot.com/2012/08/peer-feedback-good-bad-and-ugly.html

Glance, D.G. Forsey, M. & Riley, M. (2013) The Pedagogical Foundations of Massive Open Online Courses. *First Monday*, 18(5-6). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4350/3673

Haber, J. (2013) MOOCs and Grading - The Obviousity Index. *Degree of Freedom*. Retrieved from http://degreeoffreedom.org/moocs-grading-obviousity-index/

Harlen, W. (2007) Designing a fair and effective assessment system. In *Proceedings of the 2007 British Educational Research Association (BERA) Annual Conference*. September 8, 2007. London, UK.

Ho, A. D., Reich, J., Nesterko, S., Seaton, D. T., Mullaney, T., Waldo, J., & Chuang, I. (2014). HarvardX and MITx: The first year of open online courses (*HarvardX and MITx Working Paper* No. 1).

Jonsson, A., & Svingby, G. (2007) The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2007), pp. 130-144.

Kolowich, S. (2014) Writing Instructor, Skeptical of Automated Grading, Pits Machine vs. Machine. *The Chronicle of Higher Education*, April 24, 2014.

Kulkarni, C., Wei, K. P., Le, H. Chia, D., Papadopoulos, K., Cheng, J., Koller, D., and Klemmer, S.R. (2013) Peer and Self Assessment in Massive Online Classes. *ACM Transactions on Computer-Human Interaction*, 9(4).

Kulkarni, C. (2014) Making distance matter: leveraging scale and diversity in massive online classes. In *Proceedings of the 27th ACM User Interface Software and Technology (UIST) Symposium*, October 5-8, 2014, Honolulu, HI.

Larson, S. (2014) I Failed My Online Course - But Learned A Lot About Internet Education. *ReadWrite*. Retrieved from http://readwrite.com/2014/03/08/mooc-moocs-online-internet-education-fail

Lines, D. (2004) Developing a variety of assessment methods, including self and peer assessment - An overview. *Quality Assurance Agency for Higher Education - Assessment Workshop*, March 19, 2004. Retrieved from http://www.enhancementthemes.ac.uk/docs/report/developing-a-variety-of-assessment-methods-including-self-and-peer-assessment-overview.pdf?sfvrsn=18

Mandernach, J. & Garrett, J. (2014) Effective Feedback Strategies for the Online Classroom. *White Paper*. Magna Publications, Inc., Madison, WI.

Mayfield, E. (2013) Six Ways the edX Announcement Gets Automated Essay Grading Wrong. Retrieved from http://mfeldstein.com/si-ways-the-edx-announcement-gets-automated-essay-grading-wrong/

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

Morrison, S. & Free, K. (2001) Writing multiple-choice test items that promote and measure critical thinking. *Journal of Nursing Education*, 40(1).

NCTE - 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

Nielson, B. (2014) How to MOOC: Meaningful Assessment through Real-World Problem Solving. *Your Training Edge*. Retrieved from http://www.yourtrainingedge.com/how-to-mooc-meaningful-assessment-through-real-world-problem-solving-2/

Perelman, L.C. (2013) "Critique of Mark D. Shermis & Ben Hamner, 'Contrasting State-of-the-Art Automated Scoring of Essays: Analysis.'" *The Journal of Writing Assessment*, (6)1.

Reich, J. (2012) Grading Automated Essay Scoring Programs – Part 1. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/edtechresearcher/2012/04/grading_automated_essay_scoring_programs-_part_i_bjfr.html

Roland, T. (2015) *Personal interview*, February 17, 2015.

Sandeen, C. (2013) Assessment's Place in the New MOOC World. *Research & Practice in Assessment*, Volume 8. Retrieved from http://www.rpajournal.com/dev/wp-content/uploads/2013/05/SF1.pdf

Shah, N. B., Bradley, J. K., Parekh, A., Wainwright, M., & Ramchandran, K. (2013) A case for ordinal peer-evaluation in MOOCs. *NIPS Workshop on Data Driven Education*, December 10, 2013, Lake Tahoe, NV. Retrieved from http://lytics.stanford.edu/datadriveneducation/papers/shahetal.pdf

Shah, N.B., Bradley, J., Balakrishnan, S., Parekh, A., Ramchandran, K., & Wainwright, M. (2014) Some Scaling Laws for MOOC Assessments. *ACM KDD 2014 Workshop on Data Mining for Educational Assessment and Feedback*, August 24-27, 2014, New York, NY. Retrieved from http://www.stat.berkeley.edu/~sbalakri/Papers/MOOC14.pdf

Shermis, M.D. & Hammer, B. (2012) Contrasting State-of-the-Art Automated Scoring of Essays: Analysis. Paper presented at *the annual meeting of the National Council of Measurement in Education*, April 14-26, 2012, Vancouver, BC, Canada.

Solomon, E.A. (2013) MOOCs: A Review. *The Tech*, 133(3). Retrieved from http://tech.mit.edu/V133/N2/mooc.html

Vander Ark, T. (2012) Better Tests, More Writing, Deeper Learning. *Getting Smart*. Retrieved from http://gettingsmart.com/2012/04/better-tests-more-writing-deeper-learning/

Watters, A. (2012) The Problems with Peer Grading in Coursera. *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/blogs/hack-higher-education/problems-peer-grading-coursera

Watters, A. (2013) Tossing Sabots into the Automated Essay Grading Machine. *Hack Education*. Retrieved from http://hackeducation.com/2012/04/15/robot-essay-graders/

Winerip, M. (2012) Facing a Robo-Grader? Just keep Obfuscating Mellifluously. *The New York Times*, April 22, 2012.

Young, J.R. (2012) Dozens of Plagiarism Incidents Are Reported in Coursera's Free Online Courses. *The Chronicle of Higher Education*. August 16, 2012. Retrieved from http://chronicle.com/article/Dozens-of-Plagiarism-Incidents/133697/