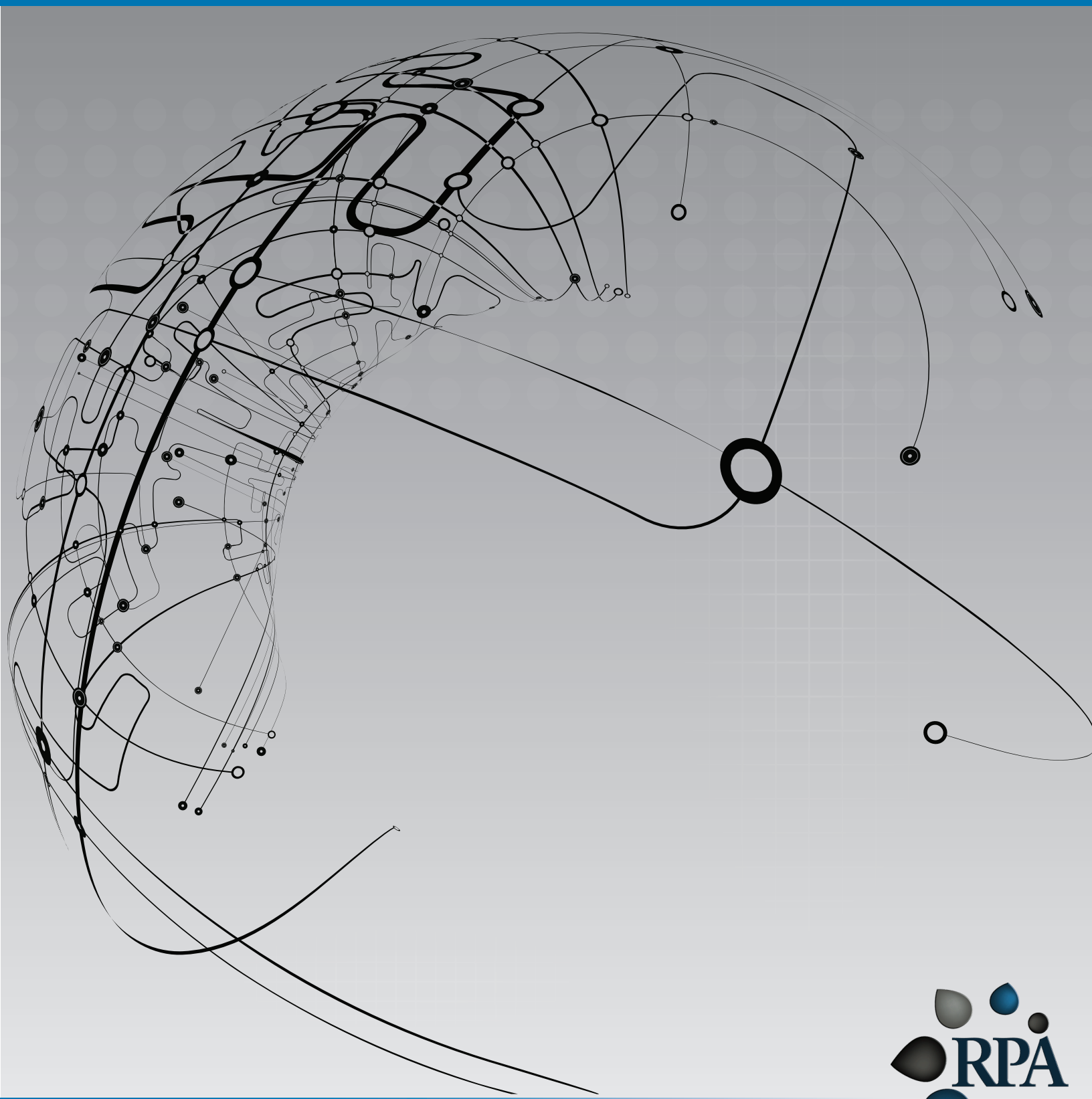
RESEARCH & PRACTICE IN ASSESSMENT

Special Issue: MOOCs & Technology

VOLUME EIGHT | SUMMER 2013
www.RPAjournal.com
ISSN # 2161-4210



RPA

A PUBLICATION OF THE VIRGINIA ASSESSMENT GROUP

RESEARCH & PRACTICE IN ASSESSMENT

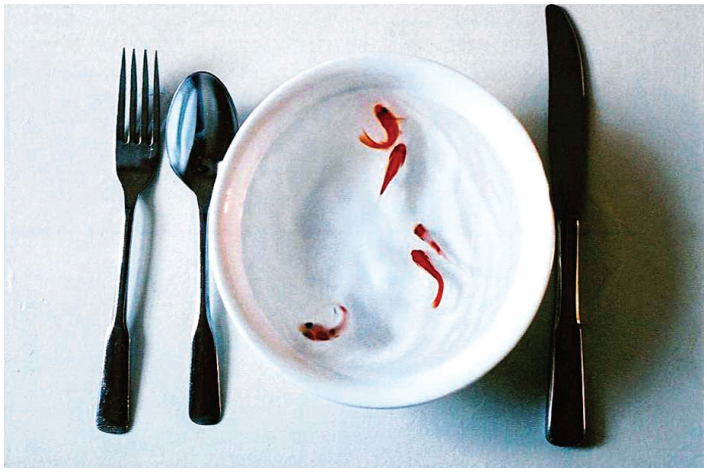
The goal of *Research & Practice in Assessment* is to serve the assessment community as an online journal focusing on higher education assessment. It is dedicated to the advancement of scholarly discussion amongst researchers and practitioners in this evolving field. The journal originated from the Board of the Virginia Assessment Group, one of the oldest continuing professional higher education assessment organizations in the United States. *Research & Practice in Assessment* is a peer-reviewed publication that uses a double-blind review process. Approximately 40% of submissions are accepted for issues that are published twice annually. *Research & Practice in Assessment* is listed in Cabell's Directory and indexed by EBSCO, Gale, and ProQuest.

TABLE OF CONTENTS

CALL FOR PAPERS

Research & Practice in Assessment is currently soliciting articles and reviews for its Winter 2013 issue. Manuscripts submitted to *RPA* may be related to various higher education assessment themes, and should adopt either an assessment measurement or an assessment policy/foundations framework. Contributions to be evaluated by the *RPA* Review Board should be submitted no later than September 1. Manuscripts must comply with the *RPA* Submission Guidelines and be sent electronically to:

editor@rpajournal.com



-Ruminate page 60

Published by:
VIRGINIA ASSESSMENT GROUP
www.virginiaassessment.org

For questions about Virginia Assessment Group membership opportunities email webmaster@virginiaassessment.org. We welcome members and non-members to join our community. If you would like to be informed of Virginia Assessment Group events, please contact us and we will add you to our distribution list.

Publication Design by Patrice Brown
www.patricebrown.net

FROM THE EDITOR

New Combinations for Higher Education

To produce means to combine materials and forces within our reach. To produce other things, or the same things by a different method, means to combine these materials and forces differently… Development [or innovation] in our sense is then defined by the carrying out of new combinations. (Joseph A. Schumpeter, Theory of Economic Development, p. 65–66)

Many mainstream publications were generous in their coverage of massive open online courses (MOOCs) as the narrative rapidly unfolded throughout 2012. While reports focused on the myriad opinions of and prophecies about this new educational context, two characteristics in these pieces were often absent: the examination of data and the application of theoretical frameworks. What was missing from the discourse was an engagement with our existing frames of knowledge. This neglected aspect is significant because the innovation inherent within MOOCs is not that *new* knowledge is being employed; rather, it is that existing knowledge is being used in new combinations. The early 20th century economist, Joseph Schumpeter purported that innovation results from new combinations of knowledge, equipment, markets and resources. The innovation is in the combination.

This distinction regarding the innovative nature of MOOCs is important because it addresses the manner in which the articles in this issue engage the MOOC context. The authors herein examine the new context using existing frames of knowledge, these include: connectivism, item response theory, research into student success and persistence, theories of online learning, calibrated peer review, and the assessment of writing, to name a few. In doing so, we are invited on the one hand to consider the extent to which MOOCs may advance our educational frameworks and knowledge. Yet, on the other hand, as Stimpson aptly articulates in her review, we are reminded to be mindful of some who may "dismiss the past in order to legitimate the brave new world that will replace it."

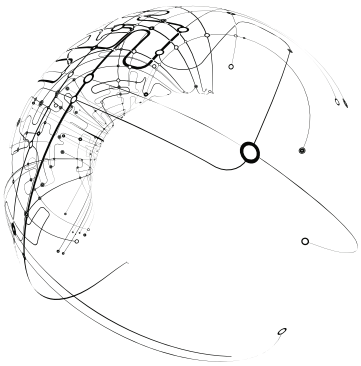
The Summer 2013 issue of *RPA* opens with an overview penned by Cathy Sandeen of the *American Council on Education* who describes the organizational distinctions between the three major MOOC providers, macro social factors driving change, and the vital role of the assessment profession in this new model of education. In a study on learning, Breslow et al. offer some of the first published empirical data from a MOOC course. The authors examine course components and how student achievement and persistence can be conceptualized in the first MOOC course offered by edX. Meyer and Zhu introduce readers to item response theory, scale linking and score equating in order to discuss the evaluation of student learning in MOOCs that yield fair and equitable test scores. Stephen Balfour, a Director of Information Technology at Texas A&M, navigates readers through the intricacies of scaling the assessment of student writing assignments by contrasting two unique technologies: automated essay scoring (AES) and calibrated peer review (CPR).

In an insightful review, Catharine Stimpson, former President of the *Modern Languages Association*, engages the recently released book by Salman Khan (of Khan's Academy) entitled, *The One World Schoolhouse.* A second review is presented in an admirable mentoring approach by a team of James Madison University graduate students and their professor who dialogue with Richard Shavelson and his work, *Measuring College Learning Responsibly.* Lastly, Katie Busby encourages assessment professionals to give serious consideration to the recent book authored by Keeling and Hersh, *We're Losing Our Minds.*

Finally, this issue ends with the Ruminate section as further consideration is given to Schumpeter's notion of new combinations in a provocative visual form. For many, I encourage you to begin the issue here, contemplating the photography of Huong Fralin and various excerpts on innovation. Art has a way of beckoning us to consider the circumstances of life through a new lens…which is the aim of this issue.

Regards,

Liberty University



Abstract

Cathy Sandeen is Vice President for Education Attainment and Innovation at American Council on Education, the nation's largest organization providing leadership and advocacy for all sectors of higher education. Within this role she develops and articulates ACE's attainment agenda and currently oversees a large research project on Massive Open Online Courses. Until January 2013, she served as dean of Continuing Education and UCLA Extension at the University of California, Los Angeles. This essay provides a brief primer on the evolution of MOOCs, an overview of major forces and trends shaping this evolution, and the role of assessment within the MOOC context.

AUTHOR

Cathy Sandeen, Ph.D.
American Council on Education

CORRESPONDENCE

Email
CSandeen@acenet.edu

Assessment's Place in the New MOOC World

When Massive Open Online Courses, or MOOCs, propelled into our awareness in the summer of 2012, they were either hailed as the solution to closing the postsecondary attainment gap in the U.S. or denounced as an extremely disruptive technology that would change higher education as we know it. Now, a year later, we realize the truth is probably somewhere in between. MOOCs certainly attracted the attention of the higher education community, they fostered a great deal of innovation, experimentation, discussion and debate, and they gave us a vision of how we might scale education with quality. For many people they also appear to have legitimized online teaching and learning, an educational practice that has existed for fifteen years or more.

Of course, the dynamics of developing a business model to finance MOOCs in a sustainable way and integrating this format into traditional degree programs are still evolving. One of the most promising aspects of MOOCs is that assessment of student learning has become central to any conversation. In this new MOOC world, assessment is not an "after-the-fact add on," but must be fully considered and integrated from the beginning of course design. In this way, MOOCs provide an opportunity for the assessment community to move to the center in one of the most exciting and potentially important conversations emerging in higher education today.

Description and History

In the online education world, the acronym, "MOOC," is not that new. It was coined in 2008 for an online course in "Connectivism and Connective Knowledge," offered by the University of Manitoba. Following the spirit of the open courseware movement, the university also opened up the course to online "auditors," students who joined the course for free. Unexpectedly, over 2,000 additional students enrolled on this basis. A movement was born ("Defining a MOOC," 2013).

Between 2008 and 2011 a number of institutions experimented with the MOOC concept continuing in the open educational resources vein. These courses are based on open resources, are free to students, have no requirements to enroll, have no enrollment limits, have relatively low levels of faculty facilitation, and encouragement community formation–but offer no academic credit. Many of these early experimental courses were developed by Canadian institutions. Not all courses achieved the extraordinarily high enrollments we see today, but many of the other components of and practices within contemporary MOOCs evolved during this time. The first recorded U.S. MOOC appeared in 2011, a course called "Online Learning Today and Tomorrow," with over 2,500 students, offered by the University of Illinois Springfield ("Defining a MOOC," 2013).

The enthusiasm behind the development of MOOC platforms is closely linked to the personal experiences of the founders.

Open online education providers such as the Khan Academy, TED, and iTunesU also emerged during this time. These providers offered high-quality educationally-oriented video content that attracted large numbers of viewers. The content was rarely organized into full courses and did not offer academic credit. Content offered by these organizations could be considered supplementary to formal coursework, such as the tutorials offered by Khan Academy. These sources also tended to appeal to individuals seeking general knowledge or enrichment rather than progress toward a degree or credential.

MOOCs entered the popular vernacular in the summer of 2012 with the rapid growth of enrollments in the three major MOOC platforms. Coursera and Udacity are two for-profit Silicon Valley, California start-ups, each led by Stanford University professors. The third, the non-profit organization, edX, led by a MIT professor, was initially a partnership between MIT and Harvard, but now is a consortium of a number of universities.

Prior to forming these entities, the faculty involved had experimented with teaching their own MOOCs. For example, Sebastian Thrun of Udacity taught introduction to artificial intelligence. Andrew Ng of Coursera taught a course in machine learning. Anant Agarwal of edX taught a course in circuits and electronics. All of these fully online courses enrolled thousands of students from around the world—and in some cases enrollments in an individual class exceeded 100,000 students. The enthusiasm behind the development of MOOC platforms is closely linked to the personal experiences of the founders.

The three major MOOC platforms are somewhat distinct from each other in terms of mission, strategy, and tactics. At the risk of over simplifying, I will attempt a brief overview. With a distinct access mission, Coursera has the largest enrollment with over 3.7 million students at the time of this writing. The firm uses a decentralized model, partnering with largely elite, "name brand," universities in the U.S. and globally (though Coursera has diversified somewhat) that are mainly responsible for delivering faculty and content. Course content leans toward upper division, specialized courses. Coursera provides the platform and various instructional and assessment tools, format guidelines, course development support, marketing, enrollment, and customer and technical support.

With a mission of fostering access and successful learning outcomes for students currently not well served by higher education, Udacity is the most vertically integrated of the three, employing a high degree of instructional design, integrated feedback and assessment tools within its courses as well as providing platform, marketing, and student support. Due mainly to their detailed and painstaking production methods, Udacity has completed fewer courses to date and tends to offer a large proportion of foundational, basic courses, especially in math and science areas.

The third MOOC platform, edX, is somewhere in between. The nonprofit start-up has formed partnerships with universities who provide content; edX also directly contributes to course and assessment design, though perhaps to a lesser degree than Udacity. Each platform collects a wealth of data on how students are interacting with their courses and the outcomes of their efforts. A number of other MOOC platforms have emerged on existing online learning management systems (e.g., Blackboard or Canvas). In these cases, the university providing the MOOC is responsible for the course design within platform parameters. The MOOCs phenomenon is not isolated to the U.S. The Open University in the UK, for example, an

institution with a deep history in distance and online education, has launched its own MOOC initiative called Futurelearn.

Emerging Issues

MOOCs were initially offered at no cost to students and on a no credit basis. The courses were open in the sense they had no prerequisites or admission requirements. Many students enrolled to "test the waters" in a new subject area or for their own personal enrichment or professional development. The majority of students who enrolled did not complete their courses. Within the "no credit context," course completions are estimated to be less than 10% (Agarwala, 2013). However, student motivations for enrolling in MOOCs vary and perhaps completion rates do not tell the whole story. Plus, in a course that enrolls 100,000 students, 10% completion is still a significant number of students.

We might note the use of the term "open" in MOOC is a bit of a misnomer. For the most part, ownership of course content and platform design is asserted and protected by course developers, therefore allowing them to monetize their intellectual property in some manner. Only edX provides open educational and platform resources in the normal sense of "open," that is material that is freely open and available for use or adaptation by others.

Earlier this year a number of formal and informal experiments and pilots emerged in an attempt to recognize and validate student learning and/or to provide "transcriptable" academic credit. For example, anecdotal reports in the media described students who completed computer science courses through MOOCs and subsequently listed them on their resumes or Linkedin profiles where potential employers might notice them. One monetization strategy of for-profit MOOC providers includes acting as an "employment agency," selling information on student performance for students who opt into the service.

Individual colleges and universities began accepting MOOCs for credit with faculty approval or completion of an assessment examination given by the university itself in order to receive credit there. The University of Helsinki, Finland, is one institution that employs this model (Kurhila, 2012). Other universities licensed MOOC content and integrated that into a campus-based course that would be eligible for credit at that particular institution. San Jose State University, California, is piloting programs with Udacity and edX MOOCs in such a hybrid format ("Frequently Asked Questions," 2013) as is Antioch University in Los Angeles ("Antioch University Becomes the First," 2013).

MOOCs provide an opportunity for the assessment community to move to the center in one of the most exciting and potentially important conversations emerging in higher education today.

For the most part, however, institutions that sponsor MOOCs do not offer their own academic credit to students who complete MOOCs at that institution. In other words, currently a non-matriculated student cannot earn University of Pennsylvania credit for completing a MOOC offered by Penn. Often a MOOC is qualitatively different from a campus course—online format notwithstanding. For courses that are exactly the same or equivalent, there appears to be a firewall of sorts between providing academic credit for paying, matriculated students at a given institution versus no credit available from the institution for the masses of nonpaying students. Protecting the integrity of their full residential campus experience appears to play a role. This "firewall" may erode over time.

Digital badges are one innovation that has emerged as a means for validating student learning whether learning occurs inside or outside the academy. The Mozilla Open Badges concept is the leading example and is modeled roughly on the iconic boy or girl scout badge. As Knight and Casilli (2012) elaborate: "A 'badge' is a symbol or indicator of accomplishment, skill, quality, or interest . . . [that has] been successfully used to set goals, motivate behaviors, represent achievements, and communicate success in many contexts" (p. 279). Digital badges can be developed by any issuer. Criteria and standards for awarding the badge as well as the characteristics and reputation of the issuing organization are made transparent under Mozilla's system (Knight & Casilli, 2012). Badges tend to acknowledge narrow and specific skills and competencies and currently are a form of alternative micro-credentialing not linked to formal academic credit as we know it.

In another experiment, the American Council on Education (ACE) began applying its long-standing course review and credit recommendation service to the MOOC environment. A pilot review of a group of five courses on the Coursera platform was completed in January 2013, with all five courses recommended for some form of academic credit.[1] Coursera students are now able to opt into the for-credit option in these courses for a fee, usually in the range of $100-150 per course. Additional pilots are underway to expand the pool of courses eligible for credit recommendations as well as a study to investigate how institutions might apply MOOC credit recommendations for their students. Like nearly everything else with this educational innovation, the credit issue is evolving, but it likely will remain important in the future acceptance of and growth of MOOCs.[2]

Drivers of Change

MOOCs have focused attention on a number of more general macro trends that have tremendous potential to disrupt and change our higher education system. Much of the conversation so far has focused on open access provided by MOOCs and the potential to educate large numbers of students for a lower per-student cost. It is still debatable whether or not MOOCs will fulfill this promise. However, MOOCs have played an important role in accelerating discussion on a number of important trends. The following section provides a brief overview.

Attainment goals. The U.S. has long prided itself on having one of the most highly educated populations in the world. Unfortunately, that is no longer the case. Over the past decade, the proportion of the population with a postsecondary degree has increased far more significantly in other advanced economies than in the U.S., particularly among young people. According to the Organisation for Economic Cooperation and Development (OECD, 2012), only 42% of young Americans ages 25-34 currently hold an associate degree or higher. Contrast that to figures for South Korea at 65%, Japan at 57%, and Canada at 56% (OECD, 2012).

Further, 63% of U.S. jobs are projected to require some level of postsecondary education by 2018 (Carnevale, Smith, & Strohl, 2010). On a social level, postsecondary attainment has always been an important means for providing social equity and economic mobility to U.S. citizens. There is a huge gap to fill. We must develop the means to provide quality education at a larger scale than ever before. Most national attainment goals speak to 60% attainment by 2025. The work ahead of us is daunting and MOOCs provide some promise.

On a social level, post-secondary attainment has always been an important means for providing social equity and economic mobility to U.S. citizens.

Cost. The cost of higher education to students and families has escalated due to reductions in traditional funding sources (Archibald & Feldman, 2012; Desrochers & Kirshstein, 2012). Students are graduating with more debt than in the past and default rates on student loans are trending upward. MOOCs are currently provided for free or at low cost to students. If a student was able to transfer some credit earned by completing MOOCs, this would decrease the student's total cost to complete a degree, certificate or credential similar to students who are able to apply transfer credit from Advanced Placement or CLEP exams. Though the price currently charged to a student end user enrolled in a MOOC is negligible, the cost to produce a MOOC is not. The University of Pennsylvania, an early Coursera adopter, estimates its cost to be $50,000 per MOOC, not including faculty time (Popp, 2013). MOOCs that incorporate a high degree of design, assessment, and analytics cost much more.

Globalization. With their global reach and estimates of 60% of enrolled students from outside the U.S., MOOCs both reflect and contribute to this trend and illustrate that the U.S. higher education system continues to attract students from around the world. MOOCs also may offer the potential for domestic students who participate to become more globally aware and culturally competent. Attainment goals are frequently linked to the need for the U.S. workforce to remain globally competitive (American Council on Education, 2011).

Competency-centered models. U.S. higher education has been—and still is—oriented toward inputs rather than outcomes. If we have the best faculty, students,

libraries, facilities, and the right amount of "seat time," the argument goes, we will have optimal student learning outcomes. Most current funding formulas are based on enrollment, not graduation.

Recently there is strong evidence of a shift. Regional accrediting bodies have begun to focus attention on outcomes in their accreditation reviews. Some states are beginning to integrate performance-based metrics into their funding formulas for public institutions. In April 2013, the U.S. Department of Education approved the eligibility of Southern New Hampshire University to receive Title IV federal financial aid for students enrolled in their new competency-based degree program (Parry, 2013), signaling a distinct willingness to move beyond the traditional credit hour measure.

With its focus on outcomes, the assessment community has inherently questioned the validity of the "inputs only" model for some time. None of this conversation is completely new. However, with their potential to collect massive amounts of student data and to integrate predictive analytics and rapid assessment and feedback systems, MOOCs may have played a role in opening up the conversation to more voices and in accelerating a reorientation from inputs to outcomes.

Technology and customization. The constant forward march of technology advancement is a consistent trend throughout all aspects of our lives. Within higher education and the MOOC environment, technological advancement intersects with cognitive learning science and will allow for a higher degree of personalization and customization of content and pedagogical methods than ever before. As Soares (2011) points out, the Carnegie Mellon Online Learning Initiative has been an early adopter in designing online environments that customize content specific to individual student needs. Currently, "smart systems" can detect areas where students are having difficulties and then direct students to additional resources or practice exercises to help them learn successfully. This capacity will only continue to be developed and refined. Analytics and feedback should play an important role in contributing to the success for the many students who require additional academic preparation in foundational subjects in order to progress toward a degree or credential.

With its focus on outcomes, the assessment community has inherently questioned the validity of the "inputs only" model for some time.

Open and ubiquitous information. Information is available, open, and free. Yes, we need to acknowledge the digital divide in this country—not everyone has access to high speed internet. Still, as long as an individual has some access, information abounds. Higher education's traditional role was as repository, organizer, and disseminator of information. That role is changing.

Disaggregation of the faculty role. Related to many of the trends above, we may be witnessing an inflection point in how faculty perform their teaching duties. For centuries, postsecondary teaching has been vertically integrated: identifying a subject area, designing a course, sourcing content, organizing content, determining learning outcomes, designing exercises and assessments, teaching the course, scoring assessments, and assigning final grades (Mendenhall, 2012). The need to increase attainment, a shift to a competency-centered approach, open access to information, as well as technological advances may focus the integrated faculty role to become one of curator of information and mentor to students, the components of teaching most valued by faculty. Are we in the position now to explore whether student learning outcomes can be improved if course design, technical content sourcing, learning technology, and assessment are "outsourced" to experts, leaving sophisticated content curation, course delivery, and personalized student mentoring to those who can do that best—faculty?

Assessment Now Front and Center

One of the more interesting and promising aspects of MOOCs is the high level of experimentation and rapid prototyping of technology-based assessment that has occurred. This has very positive implications for assessment scholars and professionals. Because of the scale of MOOCs, it would be impossible to hire enough humans to conduct all assessments required in a course. Further, the mission of several MOOC providers is to improve student learning in foundational courses, especially among first generation, low-

income students, using adaptive learning and feedback mechanisms. For these reasons, assessment methods will be hardwired into a MOOC.

Standard assessment methods are applied within MOOCs, especially in subjects that can be assessed by commonly used objective means. We also are witnessing developments in the areas of machine grading and peer grading that can be used to score writing-based assessments. Other articles in this issue will address some of these methods in more detail (Balfour, 2013).

However, with their potential to collect massive amounts of student data and to integrate predictive analytics and rapid assessment and feedback systems, MOOCs may have played a role in opening up the conversation to more voices and in accelerating a reorientation from inputs to outcomes.

The majority of MOOCs offered for credit are in STEM disciplines. It will be interesting to see new developments in large scale online assessments for classes in the humanities and the arts where multiple choice exam questions are not always the most effective or accepted assessment method.

Early response by faculty to MOOC assessment experiments in machine or peer grading shows a relatively high degree of acceptance, even at this early stage. A recent survey of MOOC faculty conducted by *The Chronicle of Higher Education* indicated that 74% of respondents used automated grading; 67.1% found the technique to be "very reliable" and 30.1% found it to be "somewhat reliable." Thirty-four percent (34%) of respondents used peer grading; 25.8% found the technique to be "very reliable" and 71% found it to be "somewhat reliable" (Kolowich, 2013).

Predictive analytics and adaptive learning, methods that permit customization of content based on student learning, have assessments embedded in them. The sheer volume of data being collected on student behavior and learning while interacting with their MOOC courses may assist the assessment community in further developing and refining techniques. Techniques developed within MOOCs will no doubt migrate into other formats and settings, including traditional online and classroom-based courses.

Related to assessment, and important considerations for those interested in granting academic credit for completing MOOC courses are issues of authentication and proctoring. In short, the academic community must be confident that the person completing the course and assessments is the same person who enrolled in the course. Authentication of identity is a common concern in standardized testing and various methods have been employed to verify identity, most commonly government issued photo identification, as well as newer biometric techniques like palm vein recognition.

Many of these methods can be converted to the MOOC environment. New methods are being developed frequently, like keystroke recognition and queries based on public record information (e.g., past addresses) that only an individual would know in detail. Authentication might be required for each quiz, assignment, or each time a student logs onto the MOOC platform.

Because of cost, proctoring typically occurs for the summative assessment only and is handled in one of two ways. One method requires the student to complete the exam at a physical testing center (a public library, educational institution, or private testing facility). This requires the student to travel to the testing center site, making it difficult for more remote students to participate.

The second method is webcam proctoring. The student is monitored throughout the time of the exam over a webcam. Proctors first scan the room via the camera and then ensure that the student is not consulting online or other resources or people for the duration of the exam. Authentication and proctoring have associated costs. Currently, these services are offered to MOOC students on an optional basis for a nominal fee (usually $100—$150) and typically would be required for a student to earn academic credit for completing the course. Authentication and proctoring are vital elements to provide a high degree of confidence in assessments within MOOCs. Expect to see many more technology solutions developed in the near future.

Final Thoughts

The rise of MOOCs is an extremely positive development for assessment scholars and practitioners. MOOCs have focused our attention and have fostered much excitement, experimentation, discussion and debate like nothing I have seen in my decades-long career in higher education. MOOCs represent a rapidly evolving landscape. I applaud *Research & Practice in Assessment (RPA)* for diving into the waters to be an early participant in this conversation. I expect MOOCs to continue to develop and evolve and I expect what we are learning in the MOOC environment to inform and to become integrated in other learning contexts.

Within the MOOC world, assessment is a central feature of design from the very beginning. In this new context, assessment is less about compliance than about supporting student learning outcomes and ultimately student success and attainment—directly in the center as it should be.

End Notes

[1] Four courses, Pre-Calculus from the University of California, Irvine; Introduction to Genetics and Evolution from Duke University; Bioelectricity: A Quantitative Approach from Duke University; Calculus: Single Variable from the University of Pennsylvania, received recommendations for undergraduate credit. Algebra from the University of California, Irvine, received a recommendation for mathematics vocational credit.

[2] In full disclosure, the author oversees this program as part of her responsibilities at ACE. The ACE Credit recommendation service has existed for over 50 years to assess and assign credit recommendations for formal learning that does not take place in a university setting (extra-institutional learning), like military service or corporate workplace education. Teams of faculty as well as pedagogical and testing experts review educational activities and provide recommendations on credit equivalencies. ACE authorizes production of student transcripts with these credit recommendations that may, at the discretion of the degree-granting institution, be applied toward a degree or other academic program. ACE has a network of 2,000 institutions that regularly consider and accept these credit recommendations.

References

Agarwala, M. (2013). A research summary of MOOC completion rates. Retrieved from http://edlab.tc.columbia.edu/index.php?q=node/8990

American Council on Education. (2011). Strength through global leadership and engagement: U.S. higher education in the 21st century. Washington, DC: Author. Retrieved from http://www.acenet.edu/news-room/Documents/2011-CIGE-BRPReport.pdf

Antioch University becomes first US institution to offer credit for MOOC learning through Coursera. (2013). Retrieved from http://www.antioch.edu/antioch-announcement/antioch-university-becomes-first-us-institution-to-offer-credit-for-mooc-learning-through-coursera/

Archibald, R. B., & Feldman, D. H. (2012). The anatomy of college tuition. Washington, DC: American Council on Education. Retrieved from http://www.acenet.edu/news-room/Documents/Anatomy-of-College-Tuition.pdf

Balfour, S. P. (2013). Assessing writing in MOOCS: Automated essay scoring and Calibrated Peer Review. *Research & Practice in Assessment, 8*(1), 40-48.

Carnevale, A. P., Smith, N., & Strohl, J. (2010). Help wanted: Projections of jobs and education requirements through 2018. Washington, DC: The Georgetown University Center for Education and the Workforce. Retrieved from http://www9.georgetown.edu/grad/gppi/hpi/cew/pdfs/HelpWanted.ExecutiveSummary.pdf

Defining a MOOC. (2013). Retrieved from http://wikieducator.org/OER_university/eduMOOC_planning_group/MOOC_comparison

Desrochers, D. M., & Kirshstein, R. J. (2012). College spending in a turbulent decade: Findings from the Delta Cost Project. Washington, DC: American Institutes for Research. Retrieved from http://www.deltacostproject.org/pdfs/Delta-Cost-College-Spending-In-A-Turbulent-Decade.pdf

Frequently asked questions (FAQ) SJSU plus. (2013). Retrieved from http://www.sjsu.edu/at/ec/sjsuplus/faq_sjsu_plus/

Knight, E., & Casilli, C. (2012). Mozilla open badges. In D.G. Oblinger (Ed.), *Game changers: Education and information technologies.* Louisville, CO: Educause.

Kolowich, S. (2013, March 21). The professors who make the MOOCs. *The Chronicle of Higher Education*, p. 1.

Kurhila, J. (2012). Studies in massive open online courses provided by other universities. Retrieved from http://www.cs.helsinki.fi/en/news/68231

Mendenhall, R. W. (2012). Western Governors University. In D.G. Oblinger (Ed.), *Game changers: Education and information technologies* (pp. 115-132). Louisville, CO: Educause.

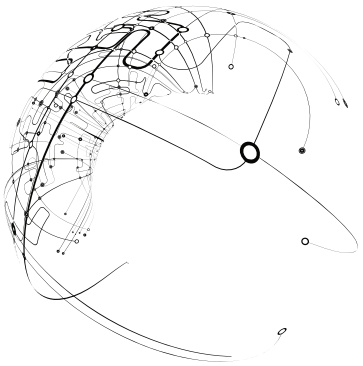
Meyer, J. P., & Zhu, S. (2013). Fair and equitable measurement of student learning in MOOCs: An introduction to item response theory, scale linking, and score equating. *Research & Practice in Assessment, 8*(1), 26-39.

Organisation for Economic Cooperation and Development (OECD). (2012). Education at a glance 2012. Washington, DC: Author.

Parry, M. (2013, April 18). Competency-based education advances with U.S. approval of program. *The Chronicle of Higher Education.* Retrieved from http://chronicle.com/blogs/wiredcampus/u-s-education-department-gives-a-boost-to-competency-based-education/43439?cid=at&utm_source=at&utm_medium=en

Popp, T. (2013, March/April). MOOC U. *The Pennsylvania Gazette*, pp. 58-63.

Soares, L. (2011). The 'personalization' of higher education: Using technology to enhance the college experience. Retrieved from http://www.americanprogress.org/issues/labor/report/2011/10/04/10484/the-personalization-of-higher-education/



Abstract

"Circuits and Electronics" (6.002x), which began in March 2012, was the first MOOC developed by edX, the consortium led by MIT and Harvard. Over 155,000 students initially registered for 6.002x, which was composed of video lectures, interactive problems, online laboratories, and a discussion forum. As the course ended in June 2012, researchers began to analyze the rich sources of data it generated. This article describes both the first stage of this research, which examined the students' use of resources by time spent on each, and a second stage that is producing an in-depth picture of who the 6.002x students were, how their own background and capabilities related to their achievement and persistence, and how their interactions with 6.002x's curricular and pedagogical components contributed to their level of success in the course.

AUTHORS

Lori Breslow, Ph.D.
Massachusetts Institute of Technology

David E. Pritchard, Ph.D.
Massachusetts Institute of Technology

Jennifer DeBoer, Ph.D.
Massachusetts Institute of Technology

Glenda S. Stump, Ph.D.
Massachusetts Institute of Technology

Andrew D. Ho, Ph.D.
Harvard University

Daniel T. Seaton, Ph.D.
Massachusetts Institute of Technology

CORRESPONDENCE

Email
lrb@mit.edu

Parts of this work have been submitted for presentation/publication at the American Educational Research Association (AERA) conference, the Educational Data Mining (EDM) conference, the MIT Learning International Networks Consortium conference (LINC), and *Communications, Association for Computer Machinery*. We are grateful for assistance with the research and presentation of the data in this paper from Dr. Y. Bergner.

Studying Learning in the Worldwide Classroom
Research into edX's First MOOC

From the launch of edX, the joint venture between MIT and Harvard to create and disseminate massive online open courses (MOOCs), the leaders of both institutions have emphasized that research into learning will be one of the initiative's core missions. As numerous articles in both the academic and popular press have pointed out, the ability of MOOCs to generate a tremendous amount of data opens up considerable opportunities for educational research. edX and Coursera, which together claim almost four and a half million enrollees, have developed platforms that track students' every click as they use instructional resources, complete assessments, and engage in social interactions. These data have the potential to help researchers identify, at a finer resolution than ever before, what contributes to students' learning and what hampers their success.

The challenge for the research and assessment communities is to determine which questions should be asked and in what priority. How can we set ourselves on a path that will produce useful short-term results while providing a foundation upon which to build? What is economically feasible? What is politically possible? How can research into MOOCs contribute to an understanding of on-campus learning? What do stakeholders—faculty, developers, government agencies, foundations, and, most importantly, students—need in order to realize the potential of digital learning, generally, and massive open online courses, specifically?

This paper describes an initial study of the data generated by MIT's first MOOC, "Circuits and Electronics" (6.002x)[1] by a team of multidisciplinary researchers from MIT and Harvard. These data include the IP addresses of all enrolled students; clickstream data that recorded each of the 230 million interactions the students had with the platform (Seaton, Bergner, Chuang, Mitros, & Pritchard, 2013); scores on homework assignments, labs, and exams; student and teaching staff posts on a discussion forum; and the results of a survey sent to the 6.002x students at the end of the course. We are trying to understand who the students were in 6.002x, how they utilized course resources, what contributed to their persistence, and what advanced or hindered their achievement. In other words, we are trying to make headway in answering the question Davidson (2012) has posited is central to on-line learning: "What modes of learning work in what situations and for whom?"

If educational researchers studying conventional brick and mortar classrooms struggle to operationalize variables like attrition and achievement, it is doubly difficult to do so for MOOCs. Participation and performance do not follow the rules by which universities have traditionally organized the teaching enterprise: MOOCs allow free and easy registration, do not require formal withdrawals, and include a large number of students who may not have any interest in completing assignments and assessments.

Our first challenge has been choosing, or in some cases adapting, the methodological approaches that can be used to analyze the data. If educational researchers studying conventional brick and mortar classrooms struggle to operationalize variables like attrition and achievement, it is doubly difficult to do so for MOOCs. Participation and performance do not follow the rules by which universities have traditionally organized the teaching enterprise: MOOCs allow free and easy registration, do not require formal withdrawals, and include a large number of students who may not have any interest in completing assignments and assessments. We are experimenting with new ways to study educational experiences in MOOCs, as naïve applications of conventional methods to the unconventional data sets they generate are likely to lead, at best, to useless results, and, at worst, to nonsensical ones.

As of this writing, our analyses have yielded a clearer picture of the first two questions we are exploring—the characteristics of the students and their use of course resources—and we report on these findings below. However, we are still in the process of developing the predictive models that will help us understand how both student background and interaction with course components contributed to or hampered the students' ability to persist in the course and, for some, to earn a certificate. Therefore, these analyses are not included in this paper.

For readers unfamiliar with MOOCs, in general, and with the MITx course, specifically, we begin with a short description of 6.002x. We then describe a first study that was carried out in summer through fall 2012, and the second stage of research that is currently underway. Finally, we consider some of the implications of our findings and suggest further directions our research, as well as other studies of MOOCs, may take.

"Circuits and Electronics" (6.002x)

"Circuits and Electronics" (6.002) is a required undergraduate course for majors in the Department of Electric Engineering and Computer Science. The first iteration of the edX version of 6.002 began in March 2012 and ran for 14 weeks through the beginning of June. It was offered again in fall 2012 and spring 2013.[2] The lead instructor for 6.002x was a MIT faculty member who has taught the on-campus version of the course over a number of years. He was joined by three other instructors, two MIT professors and edX's chief scientist, who were responsible for creating the homework assignments, labs, and tutorials, as well as five teaching assistants and three lab assistants.

Each week, a set of videos, called lecture sequences, was released. These videos, narrated by the lead instructor, average less than 10 minutes and are composed of illustrations, text, and equations drawn on a tablet (i.e., "Khan Academy" style). Interspersed among the videos are online exercises that give students an opportunity to put into practice the concepts covered in the videos. The course also includes tutorials similar to the small-group recitations that often accompany MIT lecture courses; a textbook accessible electronically; a discussion forum where students can have questions answered by other students or the teaching assistants; and a Wiki to post additional resources.

[1] 6.002x was originally introduced on MITx, the organization MIT established before it was joined by Harvard to create edX. "MITx" now identifies the specific courses developed at MIT that are distributed on the edX platform.

[2] Interested readers can access the spring 2013 version of the course at https://www.edx.org/courses/MITx/6.002x/2013_Spring/about

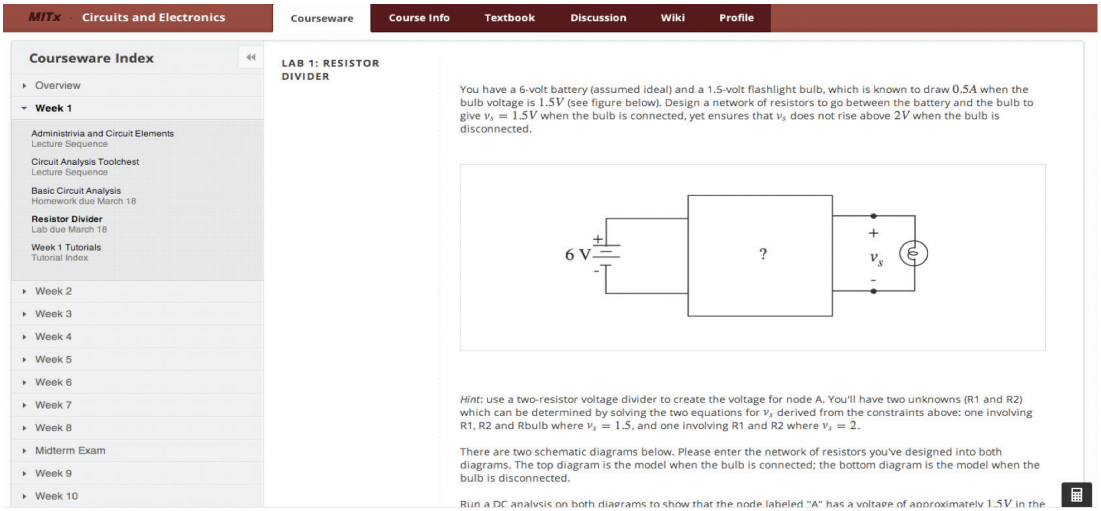


Figure 1. *Screen shot from "Circuits and Electronics" (6.002x) with navigation bar on left.*

As specified in the 6.002x syllabus, grades were based on twelve homework assignments (15%), twelve laboratory assignments (15%), a midterm (30%), and a final exam (40%). Two homework assignments and two labs could be dropped without penalty. Students needed to accrue 60 points in order to receive a certificate of completion. They received an "A" if they earned 87 points or more, a "B" for 86 through 70 points, and a "C" for 69 through 60 points. As has been widely reported, almost 155,000 people enrolled in 6.002x and just over 7,100 passed the course and earned a certificate (Hardesty, 2012).

It should be stressed that over 90% of the activity on the discussion forum resulted from students who simply viewed preexisting discussion threads, without posting questions, answers, or comments.

Within a short period of time, studies related to 6.002x were begun at MIT. During spring 2012, researchers from MIT's Research in Learning, Assessing, and Tutoring Effectively (RELATE) group began mining the data from the course to identify trends in the use of the various resources. In June 2012, MIT received an overture from the National Science Foundation to continue research on the 6.002x data set. A joint proposal was submitted by researchers from MIT's Teaching and Learning Laboratory and the Harvard Graduate School of Education to examine student demographics, online communities, and achievement and persistence among 6.002x students. As noted above, this article reports on that research to date.

First Study Explores Resource Usage

The first analysis of the 6.002x data set examined how the certificate earners allocated their time and attention over the course among the various course components. This research also explored how the behavior of certificate earners differed when solving homework versus exam problems. Each topic is addressed via the graphs below in Figure 2.

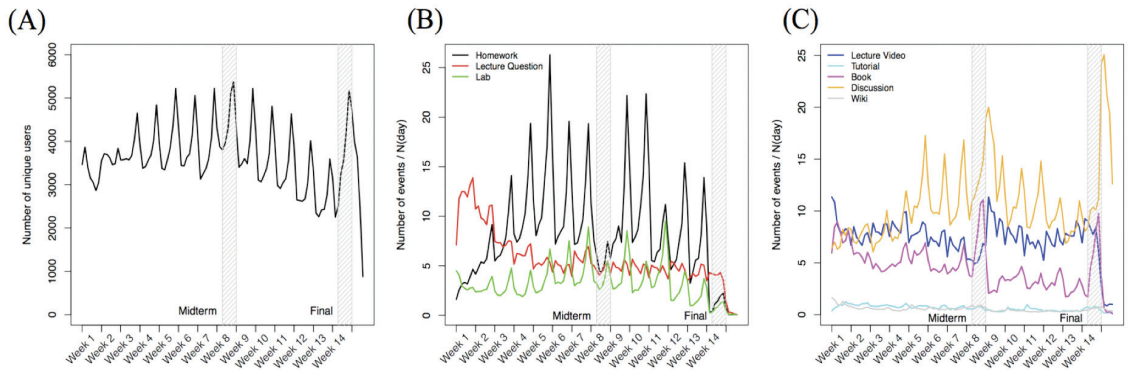


Figure 2. *Course components that were accessed in 6.002x. From left to right (A) number of unique certificate earners active per day; (B) the average number of accesses each day for assessment-based; and (C) learning-based course components.*

Plot A highlights the weekly periodicity; peaks on weekends presumably reflect both the days when spare time is available and the deadline for homework submission. In plots B and C activity is shown in hits per user each day. The three instructional resources—textbook, video lectures, and lecture questions—display little end-of-week peaking, whereas for–credit assessments (homework and labs) show marked peaks suggesting these activities were done just ahead of the deadline. The discussion forum shows similar periodicity because it is accessed while doing the homework problems (for more on the use of the discussion forum, please see below). The drop in e-text activity after the first exam is typical of textbook use that has been observed in blended on-campus courses where the textbook was a supplementary resource (that is, not part of the sequence of activities presented to students by the interface).

Students came from 194 countries, virtually all in the world. The top five countries were the United States (26,333), India (13,044), the United Kingdom (8,430), Colombia (5,900), and Spain (3,684). Although it was speculated that many Chinese students would enroll, in fact, we counted only 622 Chinese registrants.

Time represents the principal cost function for students, and it is therefore important to study how students allocated their time throughout the course. Clearly, the most time was spent on lecture videos (see Figure 3). However, the assigned work (i.e., homework and labs) took more time in toto. Use of the discussion forum was very popular considering that posting on the forum was neither for credit nor part of the main "course sequence" of prescribed activities. It should be stressed that over 90% of the activity on the discussion forum resulted from students who simply viewed preexisting discussion threads, without posting questions, answers, or comments.

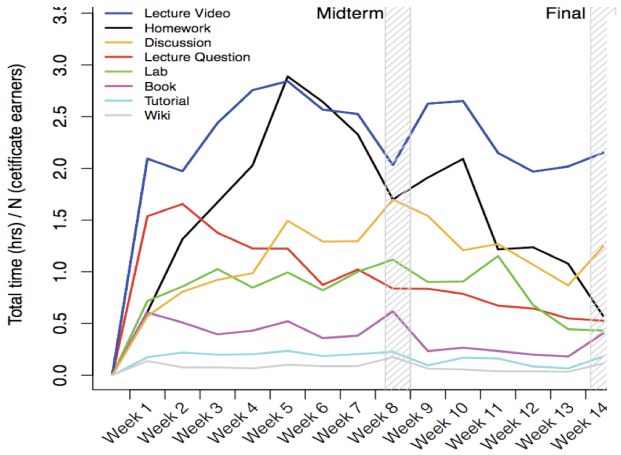


Figure 3. *Time on task. Certificate earners average time spent in hours per week on each course component. Midterm and final exam weeks are shaded.*

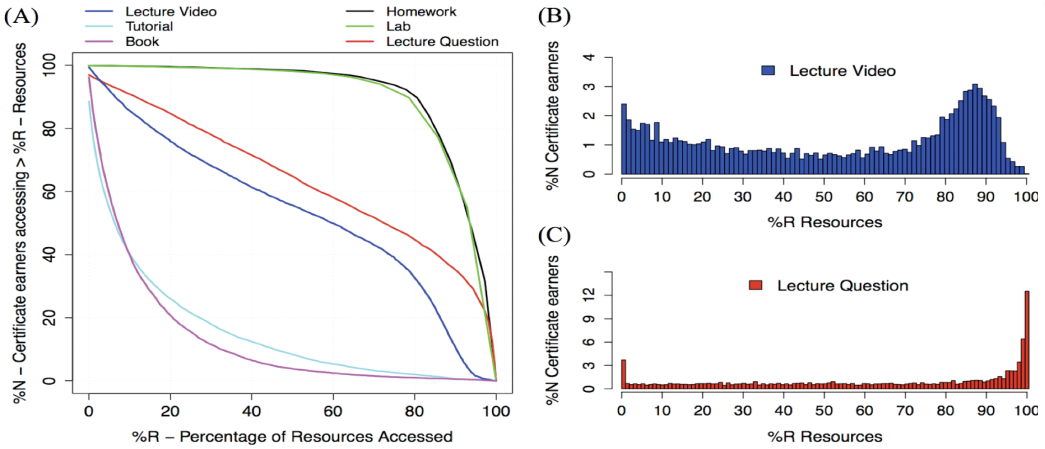


Figure 4. *Fractional use of resources: (A) the percentage of certificate earners that accessed greater than %R of that type of course resources; (B) the bimodal distribution for percentage of videos accessed; (C) distribution for the lecture questions.*

Discussions were the most frequently used resource while doing homework problems and lecture videos consumed the most time. During exams, old homework problems were most often referred to, and most time was spent with the book, which is otherwise largely neglected. This undoubtedly reflects the relevance of old homework to exams, and the ease of referencing the book for finding particular help.

Another interesting feature revealed by these data is student strategy in solving problems. By strategy, we mean which resources were most frequently consulted by the students while doing problems, and which ones were viewed for the longest time? Student strategy differs very markedly when solving homework problems versus when solving exam problems. (Note: the exams were "open course" so all resources were available to the students while they took the exams.) This finding is illustrated in Figure 5.

We know, too, from an open-ended profile edX posted at the start of the course, 67% of registrants spoke English, and 16%, the next largest group, spoke Spanish.

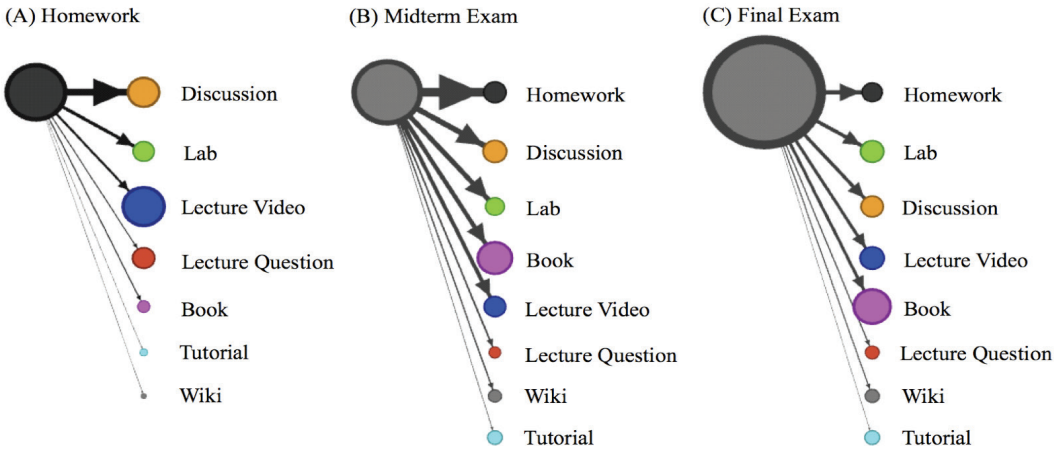


Figure 5. *Which resources are used while problem solving? Activity (hits), registered by thicker arrows, is highest for resources listed at the top. Node size represents the total time spent on that course component.*

Second Stage of Research Examines Demographics, Achievement, and Persistence

Building from the work described above, a second phase of research began in fall 2012. This study sought to answer the broad question, "Who were the students who enrolled in 6.002x, and what factors related to their level of success in the course?" This research complements the analysis of resource usage by attempting to construct a detailed picture of the 6.002x students, using multiple sampling frames: all registrants, all students who clicked on the course website, students who demonstrated different levels of engagement of the course, and certificate earners. Next, we hope to be able to identify relationships between the characteristics and capabilities of the students themselves and their success. Finally, we want to understand how the curricular and pedagogical components of 6.002x contributed to the students' ability to master the material.

Diversity in Location and Demographics

We began this research by investigating the locations from which students accessed the 6.002x site because the student's IP address was recorded each time he or she interacted with the website. We used a geolocation database to identify login locations. For nearly all IP addresses we could identify, we could determine the country from which a student logged in, and for many addresses, we could identify the city.[3] Students came from 194 countries, virtually all in the world. The top five countries were the United States (26,333), India (13,044), the United Kingdom (8,430), Colombia (5,900), and Spain (3,684). Although it was

[3] There is some error associated with this procedure, as students could log in from proxy servers or otherwise mask their IP address; however, we found less than 5% of the students were likely to be misidentified due to altered IP addresses.

speculated that many Chinese students would enroll, in fact, we counted only 622 Chinese registrants. Interestingly, we also saw a small but notable number of students who logged in from multiple countries or multiple cities within the same country. Figure 6 illustrates the widespread distribution of 6.002x students around the world.

Figure 6. *Locations of 6.002x students throughout the world.*

We know, too, from an open-ended profile edX posted at the start of the course, 67% of registrants spoke English, and 16% , the next largest group, spoke Spanish. Students who were not native English speakers formed Facebook groups to help each other with the course, and we noted a small number of posts on the discussion forum in languages other than English.

We assume some students were continuing to follow the course even if they were not doing the assignments or taking the exams.

An end-of-the-course survey was developed to gather more data about the students and their background. Because edX wanted to test the willingness of students to answer survey questions, the number of questions sent to individual students, as well as the specific questions they were asked, were distributed randomly through a link on the student's profile page. Of the 7,161 students who completed the survey, the largest group by far, 6,381 respondents, were certificate earners. However, over 800 of the respondents had not earned a certificate, so we assume some students were continuing to follow the course even if they were not doing the assignments or taking the exams. The survey questions, which were grounded in research in large-scale studies in international education, included not only demographics such as age and gender, but asked students, for example, about their home environment and their educational and professional background. This is in line with educational research (Coleman et al., 1966; Gamoran & Long, 2008) that indicates these latter variables serve as important controls in predictions of educational outcomes.

Some of the findings were not particularly surprising. For example, of the over 1,100 students who were asked about their age on the particular survey they received, most reported they were in their 20s and 30s, although the entire population of 6.002x students who responded to that question ranged from teenagers to people in their seventies. Figure 7 shows the age distribution of 6.002x students.

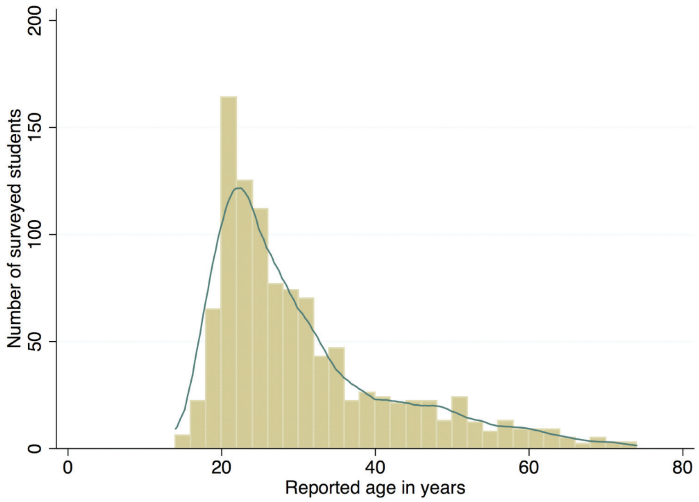


Figure 7. *Age distribution*

As might also be predicted, 88% of those who reported their gender were male. Of the survey responders who answered a question about highest degree attained, 37% had a bachelor's degree, 28% had a master's or professional degree, and 27% were high school graduates. Approximately three-quarters of those who answered the question about their background in math reported they had studied vector calculus or differential equations. In fact, the 6.002x syllabus advised students that the course required some knowledge of differential equations, along with a background in electricity and magnetism (at the level of an Advanced Placement course) and basic calculus and linear algebra.

Given that the topic of circuits and electronics has professional applications, we were not surprised to learn that over half the survey respondents reported the primary reason they enrolled in 6.002x was for the knowledge and skills they would gain. Although, interestingly, only 8.8% stated they registered for the course for "employment or job advancement opportunities." Over a quarter of the students took the course for the "personal challenge." We saw this latter motivation reflected in the discussion forum, with participants along the entire spectrum from high school students to retired electrical engineers explaining they were taking 6.002x because they wanted to see if they could "make it" through a MIT course. Figure 8 details the primary reason for enrollment for students who answered this question on the survey. There were no correlations between motivation for enrollment and success in the course. Whether students were taking 6.002x to advance their knowledge or because they wanted the challenge (we realize, of course, the two could be interrelated), it did not seem to affect their performance in the class. We are curious about how the motivation for enrollment in a course like 6.002x compares with the humanities MOOCs that have subsequently been developed.

There were no correlations between motivation for enrollment and success in the course. Whether students were taking 6.002x to advance their knowledge or because they wanted the challenge (we realize, of course, the two could be interrelated), it did not seem to affect their performance in the class.

What Contributed to Student "Success"? Predictive Modeling as the Next Step in the Analysis

The information we have collected on the students who took 6.002x offers insight into where they came from and the languages they spoke, and, for some, their educational background, the reasons they enrolled in the course, etc. Our next step is to carry out more sophisticated predictive analyses, first examining what factors individual to the students might be correlated with their success and then analyzing the relationships between the students' use of course components (e.g., hours spent doing homework, reading the textbook, or watching the lecture videos) and success. The first stage in this work is to define more precisely what we mean by "success" in a MOOC.

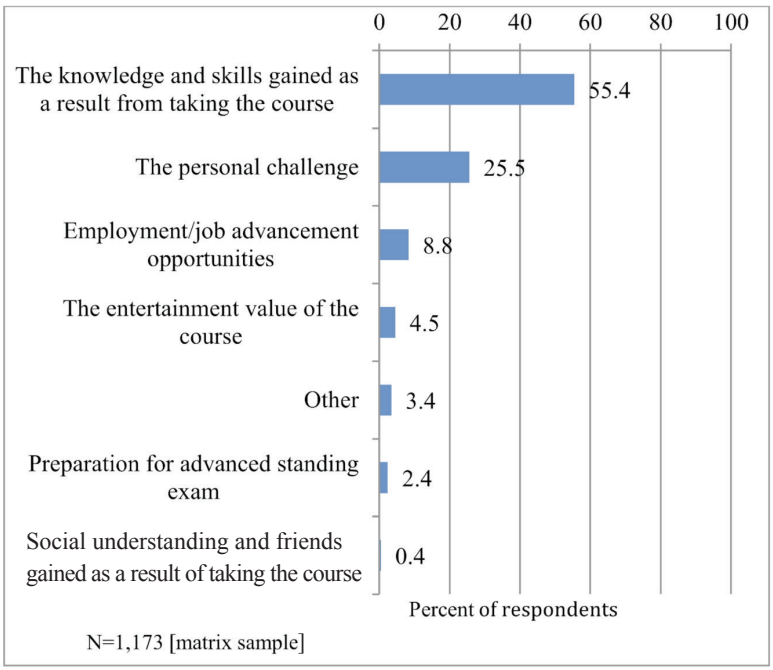


Figure 8. *Reasons for enrolling in 6.002x as reported on end-of-course survey.*

Success as Achievement

In many ways, 6.002x mirrors its on-campus counterpart: it is built from lectures, albeit shorter ones than in a traditional college course, with questions embedded between lectures so students can work with the concepts just explained in the video. 6.002x also included tutorials and laboratories. Similarly, the edX students were assessed in the same way as their on-campus counterparts—through the scores they earn on homework assignments, labs, and a midterm and final. Thus, we argue, that "success" in 6.002x can be defined as it is in the traditional college classroom, namely, by the grades students earned. We have labeled this measure of success as "achievement," and in some (but not all—please see below) of our models, "achievement" is defined as "total points in the course, weighting the individual assessments (i.e., homework, lab assignments, midterm, and final) as originally laid out in the syllabus."

Thus, we argue, that "success" in 6.002x can be defined as it is in the traditional college classroom, namely, by the grades students earned.

Using this definition, we found no relationship between age and achievement or between gender and achievement, and we found only a marginal relationship between highest degree earned and achievement. There is a correlation between students' previous course experience in mathematics and achievement, but, again, students were told at the onset of the course that they needed to know basic calculus and linear algebra, as well as have some familiarity with differential equations.

The strongest correlation we found between what we are calling "student background" and achievement was in whether or not the survey respondent "worked offline with anyone on the MITx material." The vast majority of students who answered this question (75.7%) did not. However, if a student did collaborate offline with someone else taking 6.002x, as 17.7% of the respondents reported, or with "someone who teaches or has expertise in this area," as 2.5% did, that interaction seemed to have had a beneficial effect. On average, with all other predictors being equal, a student who worked offline with someone else in the class or someone who had expertise in the subject would have a predicted score almost three points higher than someone working by him or herself. This is a noteworthy finding as it reflects what we know about on-campus instruction: that collaborating with another person, whether novice or expert, strengthens learning.

The next phase of our research is to carry out more sophisticated predictive analyses, exploring, as mentioned above, relationships between the students' use of course

components and their achievement. We want to see if certain instructional practices that are known to strengthen learning in the traditional classroom do so in MOOCs. For example, we know that mastery of knowledge and skills is often fostered by the use of "pedagogies of engagement" (e.g., Smith, Sheppard, Johnson, & Johnson, 2005), and we can explore interactive engagement in 6.002x, for instance, by estimating the impact of time spent working on online labs. Similarly, we know that retention and transfer are strengthened by practice at retrieval (e.g., Halpern & Moskel, 2003), and we can study the effect of this instructional practice by looking at the relationship between scores on practice problems and the final exam score in the course. Our goal is to begin to identify the types of curricular materials and pedagogical strategies that optimize learning outcomes for groups of learners who may differ widely in age, level of preparedness, family or work responsibilities, etc.

For some of these analyses, we have experimented with operationalizing "achievement" in two different ways: as scores on homework assignments or performance on the final. One of the features of 6.002x was that students were permitted an unlimited number of attempts at answering homework questions. Should the performance of a student who took, say, three attempts to answer a question be "equal to" the student who answered the question correctly on the first try? This is one of the issues we are grappling with. As an extension of this work, we are looking at longitudinal performance in the class. We are using panel data methods to analyze the relationship between performance on each assignment and the student's subsequent performance on the following assignment. In other words, we are taking advantage of the fine-grain resolution of the clickstream data—a weekly, daily, or even second-by-second account of student behavior and ability—to create a picture of performance over the entire class. We are also partitioning the variance in scores in a nested model, estimating the amount of variance that could be accounted for by differences between individual students and comparing that to the variance that could be explained by differences between groups of students.

On average, with all other predictors being equal, a student who worked offline with someone else in the class or someone who had expertise in the subject would have a predicted score almost three points higher than someone working by him or herself. This is a noteworthy finding as it reflects what we know about on-campus instruction: that collaborating with another person, whether novice or expert, strengthens learning.

Success as Persistence

One of the more troubling aspects of MOOCs to date is their low completion rate, which averages no more than 10%. This was true of 6.002x as well, with less than 5% of the students who signed up at any one time completing the course. Specifically, of the 154,763 students who registered for 6.002x, we know that 23,349 tried the first problem set; 10,547 made it to the mid-term; 9,318 passed the midterm; 8,240 took the final; and 7,157 earned a certificate. In other words, 6.002x was a funnel with students "leaking out" at various points along the way. Figure 9 shows the stop out rate profile for students throughout the fourteen weeks of the course.

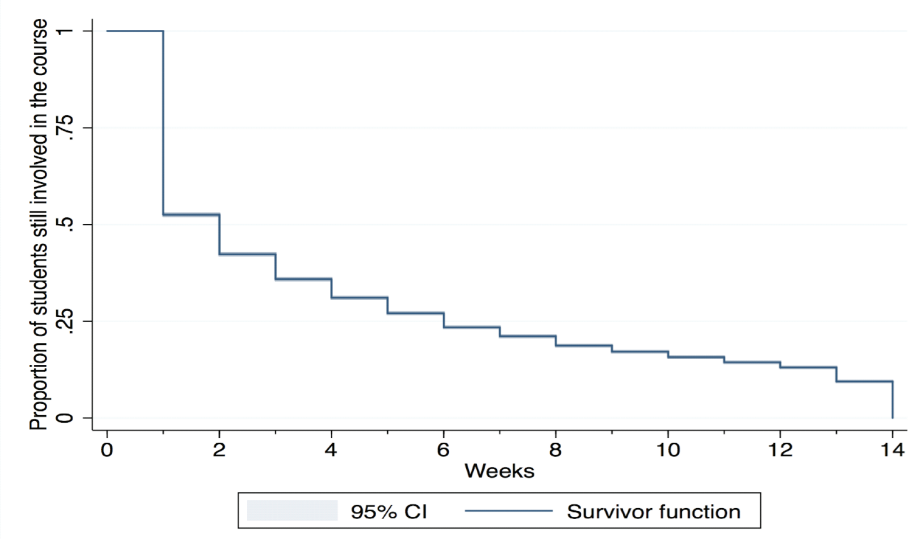


Figure 9. *Stop out rate of students throughout the course.*

We want to understand more about stop out, so we are also operationalizing "success" as "persistence throughout the duration of the course." Here, too, we are working with multiple possible definitions: persistence can be "interaction with *any part of the course* in any subsequent week" or "interaction with a *specific course component* in any subsequent week." Most investigations of students who drop out of traditional learning environments look at their trajectories over the course of a degree program or an entire academic year. Because data were collected more frequently in 6.002x, we can track users as they progressed through the course, and we can see when they chose to stop their participation.

One of the more troubling aspects of MOOCs to date is their low completion rate, which averages no more than 10%. This was true of 6.002x as well, with less than 5% of the students who signed up at any one time completing the course.

We are then estimating a survival function based on student use of resources. While the use of some resources seems to predict an increased likelihood of stopping out of the class in the next week, interactions with other resources seem to predict a decrease in likelihood of stop out. We are extending this model to look at time-varying risk functions—factors that might increase the likelihood of stopping out at the beginning of the course but have the opposite effect at the end of the course. Again, for those who completed the end-of-semester survey, we are able to control for various factors in their background.

Research on the Discussion Forum and On-Campus Use of 6.002x

The third part of this study is an in-depth look at the use of the discussion forum in 6.002x. Participation in interactive learning communities is an important instructional component of MOOCs, and investigations into the students' behavior on discussion forums may elucidate some of the possible causes of student attrition in online courses (Angelino, Williams, & Natvig, 2007; Hart, 2012). Over 12,000 discussion threads were initiated during 6.002x, including almost 100,000 individual posts, providing a rich sample for this analysis. Although the software generating the forum only allowed students to ask a question, answer a question, or make a comment, the content of the posts within those parameters was quite varied. For example, some students utilized the forum to describe how they were struggling with the material, while others offered comments that were tangential to the actual topics of the course.

However, we know that, on average, only 3% of all students participated in the discussion forum. Figure 10 below illustrates the small number of posts the vast majority of students actually made. But we know that certificate earners used the forum at a much higher rate than other students: 27.7% asked a question, 40.6% answered a question, and 36% made a comment. In total, 52% of the certificate earners were active on the forum. We are analyzing the number of comments individual students posted to see if it is predictive of that individual's level of achievement or persistence.

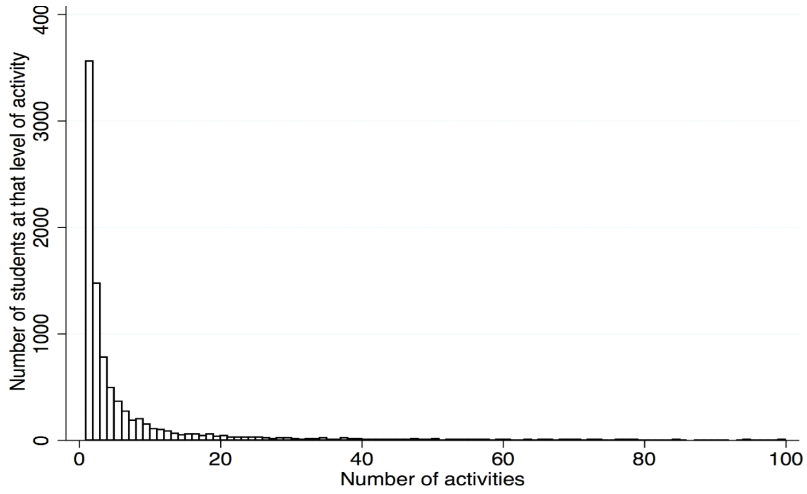


Figure 10. *Distribution of discussion board activity for students with 100 posts or less*

Our initial approach in exploring the discussion forum has been to categorize these interactions very broadly along two dimensions: (a) topic (i.e., course content, course structure or policies, course website or technology, social/affective), and (b) role of the student posting (i.e., help-seeker/ information-seeker or help-giver/information-giver). After we classify the posts using this basic schema, we will be able to describe the general purposes for which the forum was used. We hope to make a contribution to the question that has plagued those who study face-to-face collaboration, and which persists in the MOOC environment—what is the nature of the interactions that create a productive collaboration? Although previous work has shown that informal, unstructured collaboration in face-to-face educational settings is associated with higher student achievement (Stump, Hilpert, Husman, Chung, & Kim, 2011), the relationship between voluntary collaboration and achievement in the larger MOOC environment remains relatively unexplored. We want to understand how "discussion" might have helped 6.002x students to unravel a misconception, understand a difficult topic, or employ an algorithmic procedure. To do this, we are looking more specifically at threads in which students sought and received help on complex homework problems. We are examining the quantity of interactivity between question askers and responders, as well as inferences made by both parties. As yet another means of exploring these data, we are experimenting with social network analysis to see if it yields findings about the nature and longevity of group formation in 6.002x.

The last question we are exploring as part of this study is how on-campus students used 6.002x. We know that approximately 200 MIT students enrolled in 6.002x, and our data show varied levels of their participation throughout the course. We intend to interview those students who were seriously involved with 6.002x to understand their reasons for enrollment and 6.002x's impact, if any, on their studies at MIT. In addition, the Teaching and Learning Laboratory is assessing the use of materials from the edX platform in five courses being taught on campus this semester. The findings from those studies will expand our understanding of the intersection between online and on-campus educational experiences.

The optimists see MOOCs expanding access to previously disenfranchised groups of students, developing new methods of pedagogy for deeper, more sustained learning, and building global communities focused not on the latest fad or celebrity, but on education. Doomsayers predict the end of liberal learning, a generation unable to communicate in face-to-face classrooms, and even the eventual demise of the university.

Directions for Future Research

We hope our investigation of 6.002x will inform both online and on-campus teaching and learning. The appearance of MOOCs in higher education has been swift—so swift, in fact, that it could be called unprecedented. Since their introduction only a scant 18 months ago, there has been no shortage of prophecies about their potential impact. Those predictions have run the gamut from the wildly hopeful to the bleakly dire. The optimists see MOOCs expanding access to previously disenfranchised groups of students, developing new methods of pedagogy for deeper, more sustained learning, and building global communities focused not on the latest fad or celebrity, but on education. Doomsayers predict the end of liberal learning, a generation unable to communicate in face-to-face classrooms, and even the eventual demise of the university. What the two camps agree on—and what history and current events indicate—is that it is unlikely that higher educational will not be affected by MOOCs. Those effects will probably not be as dramatic as promoters or detractors would have us believe, but rather will be more nuanced and complex. A wide range of research will be needed to tease apart that impact, as well as best practices for developing and implementing MOOCs.

The authors of this paper have several areas of research they are particularly keen to explore. For example, we are interested in how the data generated by MOOCs can provide research-based comparisons of instructional strategies. A specific question, for example, is how different representations of complex concepts and phenomena (textual, graphical, mathematical) can best be used to help students master them. In general, we wish to explore how data can be utilized to provide instructors with a clearer picture of what students do or do not understand, and how that information can help them to hone their instructional skills.

Another important research question is, "How can we help students learn more per unit time?" A good way to start is to mine the logs to find what students who improve the most do—which resources they use and in which order. Then experiments will need to be done to see whether incentivizing random students helps them learn faster. The similarity of the structure of 6.002x to traditional courses means that this procedure may well permit us

to offer research-based advice to on-campus students taking a traditional course. We believe it will also be vital to better understand student motivation in an online environment. What are students' goals when they enroll in a MOOC? How do those goals relate to the interaction with various modes of instruction or course components? What facilitates or impedes their motivation to learn during a course? How can course content and its delivery support students' self-efficacy for learning? Similarly, how can online environments support students' metacognition and self-regulated learning? Do interventions such as metacognitive prompts and guided reflection improve student achievement or increase retention?

They do not follow the norms and rules that have governed university courses for centuries nor do they need to.

We are interested in policy questions, as well as the existence of MOOCs are already calling into question the nature of the university, its structure, its role in society, its accessibility to subpopulations, and its role as a mechanism for providing credentials for its students. The impact of possible certification, changes to the traditional university cost structure, and considerations of access and equity need to be understood in the new world of the MOOCs. Similarly, questions about the relationship between the social context of education beg answering.

In just the few months we have been working with the data from 6.002x, we have come to appreciate what a different animal MOOCs are, and some of the challenges they pose to researchers. The data are more numerous and at a finer grain than have ever been generated from one single course before. The students are more diverse in far more ways—in their countries of origin, the languages they speak, the prior knowledge the come to the classroom with, their age, their reasons for enrolling in the course. They do not follow the norms and rules that have governed university courses for centuries nor do they need to. Although perhaps there are not more instructional components in a MOOC than are available in the typical college course—a statement that can be contended, we suppose—those pedagogies are being used in new ways by a wider variety of people than exist in the average college classroom. All of these factors pose challenges to researchers both in framing the questions they will pursue and the methodologies they will use to answer them. But we are sure the results of research into and the assessment of MOOCs can be of value to course designers, faculty, and other teaching staff, whether they are teaching in a virtual or face-to-face classroom, and we look forward to continuing to contribute to that effort.

References

Angelino, L. M., Williams, F. K., & Natvig, D. (2007). Strategies to engage online students and reduce attrition rates. *The Journal of Educators Online, 4*(2), 1-14.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: U. S. Government Printing Office.

Davidson, C. (2012). What can MOOCs teaching us about learning? Retrieved from http://hastac.org/blogs/cathy-davidson/2012/10/01/what-can-moocs-teach-us-about-learning

Gamoran, A., & Long, D. A. (2008). *Equality of educational opportunity: A 40 year perspective.* In R. Teese, S. Lamb, & M. Duru-Bellats (Eds.), *Education and equity. Vol. 1: International perspectives on theory and policy* (Chapter 1). New York, NY: Springer Press.

Halpern, D. F., & Hakel, M. D. (2003). Applying the science of learning to the university and beyond. *Change*, 37-41.

Hardesty, L. (2012, July 16). Lessons learned from MITs's prototype course. *MIT News*. Retrieved from http://web.mit.edu/newsoffice/2012/mitx-edx-first-course-recap-0716.html

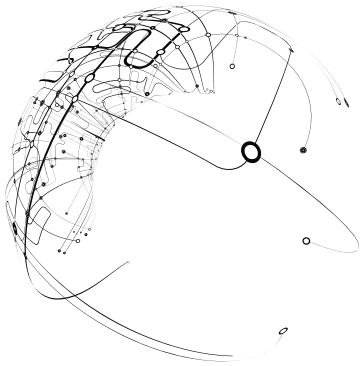
Hart, C. (2012). Factors associated with student persistence in an online program of study: A review of the literature. *Journal of Interactive Online Learning, 11*(1), 19-42.

Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P., & Pritchard, D. E. (2013). Who does what in a massive open online course? *Communications of the ACM.*

Smith, K. A., Sheppard, S. D., Johnson, D. W., & Johnson, R. T. (2005). Pedagogies of engagement: Classroom based practices. *Journal of Engineering Education, 94*(1), 87-101.

Stump, G. S., Hilpert, J., Husman, J., Chung, W. T., & Kim, W. (2011). Collaborative learning in engineering students: Gender and achievement. *Journal of Engineering Education, 100*(3), 475-497.

Lori Breslow is Director of the MIT Teaching and Learning Laboratory (TLL). David Pritchard is a Professor of Physics at MIT and Principal Investigator of the Research in Learning, Assessing and Tutoring Effectively (RELATE) group at MIT. Jennifer DeBoer is a Postdoctoral Associate in Education Research at TLL. Glenda Stump is an Associate Director for Assessment and Evaluation at TLL. Andrew Ho is an Associate Professor at the Harvard Graduate School of Education. At the time of this research, Daniel Seaton was a Postdoctoral Fellow in RELATE; he is now a Postdoctoral Associate in MIT's Office of Digital Learning (ODL). This research was supported by NSF Grant # DRL-1258448.



Abstract

Massive open online courses (MOOCs) are playing an increasingly important role in higher education around the world, but despite their popularity, the measurement of student learning in these courses is hampered by cheating and other problems that lead to unfair evaluation of student learning. In this paper, we describe a framework for maintaining test security and preventing one form of cheating in online assessments. We also introduce readers to item response theory, scale linking, and score equating to demonstrate the way these methods can produce fair and equitable test scores. Patrick Meyer is an Assistant Professor in the Curry School of Education at the University of Virginia. He is the inventor of jMetrik, an open source psychometric software program. Shi Zhu is a doctoral student in the Research, Statistics, and Evaluation program in the Curry School of Education. He holds a Ph.D. in History from Nanjing University in China.

AUTHORS

J. Patrick Meyer, Ph.D.
University of Virginia

Shi Zhu, Ph.D.
University of Virginia

CORRESPONDENCE

Email
meyerjp@virginia.edu

Fair and Equitable Measurement of Student Learning in MOOCs: An Introduction to Item Response Theory, Scale Linking, and Score Equating

The last couple of years have witnessed booming development of massive open online courses (MOOCs). These free online courses provide an innovative way of teaching and learning and make higher education accessible to a global audience. Anyone with an internet connection can take courses from top universities in the United States, Canada, Mexico, Europe, Asia, and Australia (Lewin, 2013). MOOCs hold the promise of distributing high quality courses to a global audience and making higher education accessible to people who could not otherwise afford it. Children from working-class families or low-SES backgrounds who could not attend elite universities due to economic reasons are now able to get access to these universities' teaching resources without financial difficulty. Even middle class families can look to MOOCs as a way to offset high tuition rates (Thrift, 2013). Despite the promise of MOOCs, few colleges and universities offer full course credit to students completing a MOOC. Indeed, only five of Coursea's courses are approved for course credit by the American Council on Education (Lederman, 2013), and many professors teaching MOOCs feel that students do not deserve course credit for completing a MOOC (Kolowich, 2013). The concern for course credit not only centers around course quality but also the assessment of student learning.

Online assessments are becoming more important in higher education because students who take online courses do not have many chances to communicate with their instructors and demonstrate mastery of course content in a direct way (Rovai, 2000). One obvious advantage of online assessment over a traditional test is that it can be carried out flexibly in different locations and at different time periods, and can be integrated into the online learning environment (Reeves, 2000). These assessments may simply be online versions of paper-and-pencil tests given in a traditional classroom or they may be innovative

assessments that take full advantage of resources available in computer based testing. For example, personalized e-learning systems based upon item response theory (IRT) can provide adaptive online assessments (Baylari & Montazer, 2009; Chen, Lee, & Chen, 2004) that tailor testing and course content to individual student ability. These adaptive online assessments start with items of moderate difficulty, and then change item difficulty according to a test taker's performance. Given that examinees complete different items and different numbers of items, the final score is not based upon the number of answers he or she got correct but the difficulty and discrimination levels of correctly answered questions (Challis, 2005). Once the score is known, course content can then be tailored to each individual student (see Baylari & Montazer, 2009).

Despite the innovative possibilities with online assessment, there are still some problems that cause concern among educators, policy makers, and test designers such as content disclosure, violations of intellectual property rights, system integrity (Challis, 2005), and identity security (Rovai, 2000). Perhaps the most serious threat to online assessments is cheating, a problem that has long existed in testing.

Despite the innovative possibilities with online assessment, there are still some problems that cause concern among educators, policy makers, and test designers such as content disclosure, violations of intellectual property rights, system integrity, and identity security.

Cizek (1999, 2003) identifies three types of cheating: (a) cheating by giving, taking, or receiving information from others; (b) cheating through use of prohibited materials; and (c) cheating by thwarting the testing process. These types of cheating are observed in traditional paper and pencil testing as well as online testing. Examples of cheating in an online environment include online communication, telecommunication, internet surfing (Rogers, 2006), copying and pasting from online sources (Underwood & Szabo, 2003), obtaining answer keys in an illegitimate way, taking the same assessment several times, and getting unauthorized help during the assessment (Rowe, 2008). Cheating gives dishonest examinees an unfair advantage in the assessment process and it leads assessment professionals to the wrong decision about examinees.

Cohen and Wollack (2006) describe three types of countermeasures that can be used to combat cheating and level the playing field. Human countermeasures require a proctored testing environment and entail any observational methods a test proctor can use to detect cheating. Examples include looking for a student who is conspicuously nervous or who makes frequent trips to the restroom. Electronic countermeasures are similar and may also require a formal testing environment. However, electronic countermeasures make use of technology to prevent and detect cheating. For example, a test environment may use cameras instead of a human proctor to monitor examinees or it may use special equipment to scramble cell phone signals during a test. Electronic countermeasures for an online exam may include installation of security software and IP tracking (Rogers, 2006; Rowe, 2008). Finally, psychometric countermeasures include statistical methods for the prevention and detection of cheating.

Among psychometric counter measures are procedures to limit item exposure (Cohen & Wollack, 2006). If thousands of examinees all complete the same test form, then everyone sees the same items and the risk of an examinee copying and sharing test items with others greatly increases. Prior knowledge of test items will undoubtedly give an advantage to examinees with this information and lead to a breach of standardization procedures and a lack of fairness (Cook & Eignor, 1991). One simple method for reducing item exposure and reducing the impact of cheating is the use of multiple test forms (Cizek, 1999, 2003; Cohen & Wollack, 2006; Cook & Eignor, 1991). This practice reduces exposure and it lessens the possibility that an examinee will cheat because the examinee will not know if the items for which he or she has prior knowledge will actually be on the test he or she is given. Item exposure decreases as the number of test forms increases. In an extreme case, randomly selecting items from a large item pool could result in every examinee completing a unique test form (see Lederman, 2013; Rowe, 2013).

Cook and Eignor (1991) noted that testing must be "fair and equitable" (p. 191). Use of multiple test forms improves fairness by reducing the occurrence of cheating, but it can result in inequities if one test form is easier than another. Students receiving the easier form will perform better and have a greater advantage in seeking MOOC course credit than a student who receives the more difficult form. To achieve Cook and Eignor's fair and equitable criterion, multiple test forms must be placed on a common scale. Scale linking and score

equating results in comparability among scores from different test forms and a situation in which examinees can feel indifferent about the test form they are given. The remainder of this paper discusses the use of item response theory to link and equate multiple test forms. Our discussion focuses on two test forms but it easily extends to any number of test forms or even an entire item bank. As described below, the basic steps in this framework are to: (a) collect data in a way suitable for linking, (b) estimate item and person parameters, (c) link estimates to a common scale, and (d) equate test scores to adjust for test difficulty. The first three steps are required whenever there are multiple test forms. The third step is only needed if the reporting metric is based on the observed score and not the item response theory ability score. Our aim is to introduce readers to this framework. To this end, we have omitted many of the details needed to fully implement this framework.[1]

Item Response Theory

Instructors implicitly rely on classical methods for test scaling and analysis when they create an exam or quiz score by summing the number of items answered correctly by a student. These methods are easy to implement in a classroom setting and provide for well-established methods of analyzing data and evaluating test quality. Tests designed with classical methods give instructors confidence that student scores would not change much if they had given them a different test built to the same content specifications.

Given that examinees complete different items and different numbers of items, the final score is not based upon the number of answers he or she got correct but the difficulty and discrimination levels of correctly answered questions.

Item analysis lies at the heart of evaluating the quality of tests developed through classical methods. Item difficulty and discrimination are two statistics in an item analysis. Item difficulty is the mean item score and item discrimination is the correlation between the item score and test score. These statistics allow instructors to identify problematic items such as those that are too easy or too difficult for students and items that are unrelated to the overall score. Instructors can then improve the measure by revising or eliminating poorly functioning items. An end goal of item analysis is to identify good items and maximize score reliability.

Although classical methods are widely used and easy to implement, they suffer from a number of limitations that are less evident to instructors. One limitation is that classical test theory applies to test scores, not item scores. Item difficulty and discrimination in the classical model are ad hoc statistics that guide test development. They are not parameters in the model. Through rules-of-thumb established through research and practice (see Allen & Yen, 1979), these statistics aid item selection and help optimize reliability. However, they do not quantify the contribution of an individual item to our understanding of the measured trait.

A second limitation to the classical approach is that item statistics and test characteristics are population dependent. Item difficulty will be large (i.e., easier) if a test is given to a group of gifted students, but it will be small (i.e., harder) if the same item is given to a group of academically challenged students. Population effects on item difficulty make it difficult to evaluate item quality because the statistic also reflects examinee quality. Score reliability also depends on the examinee population. It is defined as the ratio of true score variance to observed score variance. As such, scores from a population that is heterogeneous with respect to the measured trait will be more reliable than scores from a population that is homogenous. This result means that an instructor's confidence in the reproducibility of test scores depends on the group of students taking the test (Hambleton & Swaminathan, 1985).

The dependence between item and person characteristics in the classical approach also plays out at the test score level. A test will seem easy if given to a group of gifted students because the average test score will be higher than it is for the general population. Even if multiple test forms are developed to the same specifications and have similar levels of reliability, they will slightly differ in difficulty because of differences in groups taking each form. Equating must be conducted to adjust for these differences and produce comparable scores. Linear and equipercentile equating (see Kolen & Brennan, 2004) are two classical approaches to test equating that use the observed score as the basis of equating.

[1] Readers can find detailed information about test equating in Kolen and Brennan's (2004) *Test equating, scaling and linking: Methods and practices.*

Item response theory (IRT) overcomes these limitations of the classical model. IRT item statistics are estimates of parameters in the model and they can tell us about the contribution of each item to our understanding of the latent trait. Moreover, parameters are invariant to changes in the population, up to a linear transformation (Rupp & Zumbo, 2006). This statement means that if the model fits the data, item parameters will be the same in every population subject to a linear transformation. It also means that person parameters (i.e., the latent trait) will be the same in every group of items that conform to the test specifications (Bond & Fox, 2007). That is, we can obtain the same person ability estimate, within measurement error, from any set of test items. All that we need to do is apply a linear transformation to the parameters from one form to place it on the scale of another. Overcoming the limitations of classical methods does not come without a cost. At a theoretical level, IRT requires more strict assumptions and, at a practical level, it requires more training and specialized software.

Binary Item Response Models

Item difficulty and discrimination in the classical model are ad hoc statistics that guide test development.

Item response models exist for binary scored (e.g., multiple-choice) and polytomous scored (e.g., constructed response, Likert scales) test questions. For brevity, we will focus on the common unidimensional models for binary items. The most general model is the three parameter logistic (3PL) model. It has one parameter for examinee ability and three parameters for item characteristics. The model is given by

$$P(\theta) = c + (1-c)\frac{\exp[a(\theta - b)]}{1+\exp[a(\theta - b)]}.$$

The Greek letter theta, θ, is the examinee ability parameter. It represents a person's latent trait value. The exponential function is indicated by exp in this equation, and the letters *a*, *b,* and *c* represent item parameters.

Item discrimination, the *a* parameter, is the slope of the line tangent to the item characteristic curve (ICC; see Figure 1) at the point of inflection. It reflects the relationship between an item response and the latent trait. It is similar to a factor loading in factor analysis. Item discrimination is always positive. Large item discrimination values will produce an ICC with a steep curve and small values will produce a flat curve. Item difficulty, the *b* parameter, affects the location of the curve. Small difficulty values shift the whole curve to the left and large values shift it to the right. Interpretation of item difficulty in IRT is opposite that for the classical item difficulty statistic, but it is in a more intuitive direction. Small values of item difficulty are easy items, whereas large values are difficult ones. Finally, the guessing parameter, the *c* parameter, indicates the lower asymptote of the ICC. This means that an examinee with an extremely low ability level still has a small chance of answering the item correctly. It is presumed that this small chance is due to guessing on a multiple-choice test.

In the 3PL model, discrimination, difficulty, and guessing can be different for every item. Constraining these parameters leads to different IRT models. The two parameter logistic (2PL) and 1 parameter logistic (1PL) models are special cases of the 3PL. In the 2PL, the guessing parameter is fixed to zero meaning that low ability examinees have a near zero chance of answering the item correctly. The only parameters estimated in the 2PL are item discrimination and difficulty. In the 1PL model, guessing is fixed to zero and discrimination is fixed to be the same for every item but difficulty is freely estimated for every item. That is, discrimination is estimated in the 1PL but a single discrimination value is applied to all items. Item difficulty is also estimated in the 1PL but it is allowed to be different for every item. Finally, the Rasch model is a special version of the 1PL that requires the discrimination parameter to be fixed to a value of one for every item. Only the difficulty parameter is estimated in the Rasch model.

Table 1 lists item parameters for two test forms, Form X and Form Y. However, item parameters are best explained through a graph. An ICC illustrates the probability of a correct answer, $P(\theta)$, for different levels of examinee ability. Figure 1 shows the ICCs for Items 21 and 23 on Form X. As ability increases along the x-axis, the curves increase indicating that the probability of a correct answer increases as the value of the latent trait increases. Item

parameters affect the probability of a correct response and look of the ICC. Item 21 is less discriminating and difficult but involves more guessing than Item 23 (see Table 1). Because of these differences in parameters, the ICC for Item 21 is less steep, shifted to the left, and has a larger lower asymptote than Item 23.

Table 1
Item Parameters for Form X and Form Y before Linking

	Form X Item Parameters			Form Y Item Parameters		
Item	*a*	*b*	*c*	*a*	*b*	*c*
1	**1.17**	**0.56**	**0.11**	**1.31**	**1.09**	**0.10**
2	1.12	-1.39	0.24	1.25	0.35	0.22
3	0.88	-2.40	0.25	1.50	1.48	0.02
4	1.08	-2.87	0.24	1.44	-0.78	0.29
5	**0.95**	**-0.90**	**0.19**	**1.09**	**-0.30**	**0.23**
6	1.01	-0.23	0.19	1.34	-0.30	0.23
7	1.04	1.14	0.02	0.96	0.15	0.25
8	1.15	0.16	0.07	1.22	-0.62	0.22
9	**0.90**	**-0.85**	**0.20**	**1.14**	**-0.01**	**0.26**
10	1.04	0.40	0.23	1.28	-0.44	0.12
11	0.97	-0.24	0.30	1.28	-0.01	0.23
12	1.28	1.16	0.23	1.19	-0.86	0.14
13	**1.07**	**-0.39**	**0.11**	**1.22**	**0.13**	**0.06**
14	1.18	-0.23	0.13	1.26	0.07	0.21
15	0.97	-1.69	0.25	1.39	0.18	0.18
16	1.06	0.94	0.25	1.36	-0.23	0.10
17	**1.27**	**-1.19**	**0.33**	**1.53**	**-0.59**	**0.29**
18	1.29	0.81	0.16	1.06	0.37	0.06
19	0.88	0.90	0.22	0.90	0.46	0.17
20	0.94	-0.33	0.04	1.14	1.46	0.12
21	**0.77**	**-1.26**	**0.19**	**0.95**	**-0.43**	**0.26**
22	0.93	-2.29	0.23	1.08	1.20	0.07
23	1.28	0.25	0.09	1.18	-1.00	0.30
24	1.04	-3.22	0.23	0.98	0.10	0.05
25	**0.96**	**-0.66**	**0.10**	**1.14**	**-0.07**	**0.11**
26	0.93	-1.25	0.23	1.10	0.50	0.10
27	0.98	0.42	0.22	1.04	0.61	0.26
28	0.99	-0.41	0.17	1.10	0.15	0.14
29	**1.14**	**-0.51**	**0.15**	**1.36**	**-0.02**	**0.10**
30	1.00	-0.48	0.26	0.97	-1.07	0.24
Mean	1.04	-0.53	0.19	1.19	0.05	0.17
S.D.	0.13	1.16	0.08	0.17	0.67	0.08

Note: Bold font indicates common items.

Item response theory (IRT) overcomes these limitations of the classical model. IRT item statistics are estimates of parameters in the model and they can tell us about the contribution of each item to our understanding of the latent trait.

Item characteristics in IRT relate directly to test characteristics. A test characteristic curve (TCC) is the sum of all ICCs. It describes the regression of true scores on the latent trait. That is, the x-axis represents person ability, and the y-axis represents true scores. Figure 2 illustrates the TCC for Form X. It looks similar to an ICC but the y-axis is different. The y-axis ranges from the sum of the guessing parameters (5.6 in Figure 2) to the maximum possible sum score (30 in Figure 2). Because of the relationship between a TCC and an ICC, we can select items for a test in a way that achieves a desired TCC.

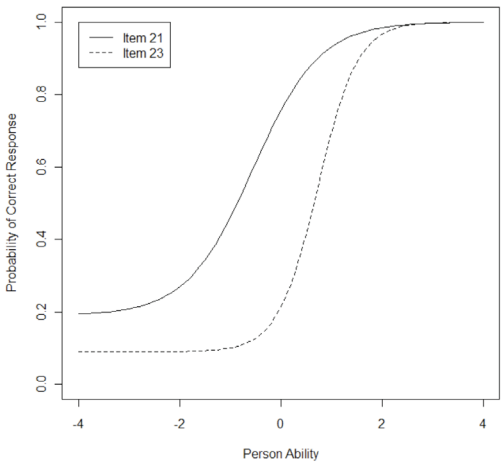


Figure 1. *Item characteristic curves for two items.*

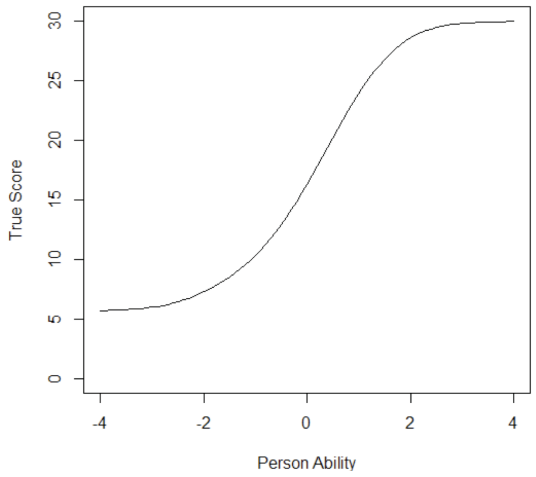


Figure 2. *Test characteristic curve for Form X.*

Another useful function in IRT is the item information function, $I_i(\theta)$. In the 3PL model it is $I_i(\theta) = \left\{a^2[1-P(\theta)]\right\}P^{-1}(\theta)\left\{[P(\theta)-c]^2/[1-c]^2\right\}$. The item information function tells us about the contribution of a single item to our understanding of the latent trait. In the Rasch and 2PL model, information is largest at the place where item difficulty equals examinee ability. Like the ICC, the difficulty parameter affects how far left or right the information curve is shifted and the item discrimination parameter affects how peaked the curve appears. Low difficulty values place information along low levels of the ability scale, whereas larger difficulty values place information at high points of the scale. In a similar vein, large discrimination values concentrate a lot of information over small range of ability levels, but small discrimination values spread a small amount of information over a wide range of the scale. That is, items with large discrimination values tell us more about the latent trait at a particular point than do items with low discrimination. Finally, as the guessing parameter increases, the amount of information decreases.

Figure 3 illustrates the effect of item parameters on the item information function. This figure involves the same two items as Figure 1. The more discriminating item (Item 23) has a more peaked information function than Item 21. It is also shifted to the right because it has a larger difficulty value than Item 21. Notice that these two items tell us very little about examinee ability values less than -2. Most of the information is concentrated between -2.0 and 2.5. To improve information at low ability levels, we should add easier items to the test (e.g., those with a difficulty less than -2.0).

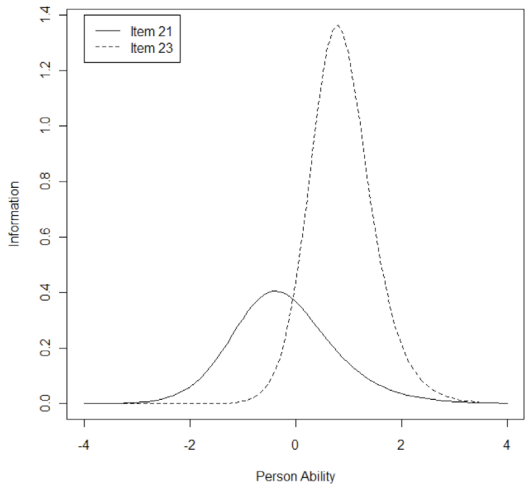


Figure 3. *Item information functions for two items.*

Information also plays a role at the test level. The test information function is the sum of all item information functions. The greater the information, the more we know about the latent trait. Consequently, we can create a test information function that targets specific ability levels, such as the passing score, by selecting items that provide a lot of information at that point. The relationship between item information functions and the test information function make evident the contribution of each item to our understanding of the latent trait. Indeed, information functions are central to many item selection routines in computerized adaptive testing (see Wainer et al., 2000).

Test information is a concept in IRT that replaces the idea of reliability from the classical model in that we aim to maximize information. The reason for maximizing information is because information is inversely related to the standard error of estimating examinee ability, $SE(\theta) = 1/\sqrt{I(\theta)}$. The ability levels with the most information are the ones that have the highest amount of measurement precision.

Parameter Estimation and Software

Marginal maximum likelihood estimation (MMLE) is a method used to obtain parameter estimates in the 2PL and 3PL models. Conditional maximum likelihood (CMLE) and joint maximum likelihood (JMLE) are alternative methods of estimation typically applied to the Rasch family of item response models. For the models discussed in this paper, all of these methods assume we are measuring a single latent trait (unidimensionality) and that items are independent at a given value of the latent trait (conditional independence; see Hambleton & Swaminathan, 1985). We will not discuss the details of these estimation methods, but on a practical level, these methods are synonymous with different types of IRT software. Programs such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), MULTILOG (Thissen, 1991), ICL (Hanson, 2002), and PARSCALE (Muraki & Bock, 1997) offer MMLE for 2PL, 3PL, and polytomous response models. WINSTEPS (Linacre, 2011) and jMetrik (Meyer, 2013) provide JMLE for Rasch family models, and the eRM (Mair & Hatzinger, 2007) package in R provides CML for Rasch family models.

We can create a test information function that targets specific ability levels, such as the passing score, by selecting items that provide a lot of information at that point.

Sample size requirements are another practical consideration for IRT. As a rule of thumb, the more parameters in the model, the larger the sample size that is needed to obtain stable parameter estimates. Rasch models require as little as 100 examinees (Wang & Chen, 2005), but the 3PL model may require at least 1,500 (Mislevy & Stocking, 1989). These sample size requirements are prohibitive for small classrooms and they are one reason why IRT is not used very often in traditional course settings. MOOCs, on the other hand, enroll tens of thousands of students, which is more than enough to obtain accurate estimates with any IRT model. Large class sizes are one reason why IRT and MOOCs are the perfect marriage.

Scale Linking in Item Response Theory

Data must be collected in a particular way in order to implement scale linking. In an equivalent groups design, each test form is given to a random sample of examinees. Items can be completely unique to each test form because the groups are randomly equivalent; test forms are considered to be the only reason for difference in test performance. Consequently, person ability estimates form the basis of scale transformation coefficients that place each form on a common scale.

A popular alternative to the equivalent groups design is the common item nonequivalent groups design (Kolen & Brennan, 1987). In this design, two different groups receive a different test form. For example, one group receives Form X and another group receives Form Y. Each test form includes a set of items unique to the form and a set of items common to both forms. Examinees are considered to be the only reason for differences in performance and parameter estimates for the common items form the basis of scale transformation coefficients. This design is easy to implement in practice but it requires great care in creating the set of common items that are embedded on each test form.

Overall, each form is designed to measure the same content and adhere to the same test specifications. Common items embedded in each form are selected to be a mini or midi version of the complete test and they are placed in about the same position on each form (Kolen & Brennan, 2004; Liu, Sinharay, Holland, Curley, & Feigenbaum, 2011). In a mini version of the test, common items cover the same range of difficulty values as the complete test, and in a midi version, common items cover a narrower range of difficulty. Table 1 demonstrates a common item design with item parameters from two different forms. The items in bold are the items shared by both forms. Once we collect data we can estimate parameters and place both forms on the same scale.

As noted earlier, parameters in an IRT model are invariant up to a linear transformation. If you apply a linear transformation to the person ability parameter and the same transformation to the item parameters, the probability of a correct response remains the same as it was prior to any transformation. This implies that there are no unique parameter values that determine the scale; any linear transformation of the parameters would result in the same probabilities. This problem is referred to as scale indeterminacy and it is resolved in practice by arbitrarily setting the person ability scale to have a mean of zero and a standard deviation of one during

the estimation process. A consequence of resolving scale indeterminacy in this way is that an item that is included on two different test forms will have different parameter estimates. However, we can use the differences in item parameter estimates from both forms to identify the linear transformation that places both forms on the same scale.

Steps for linking test forms to a common scale differ depending on whether estimation is conducted concurrently or separately. In concurrent calibration, data from all test forms are combined into a single data set and the parameters are estimated simultaneously. The overlap in common items will result in estimates that are on a common scale. No further work is needed to place Form X parameters on the scale of Form Y. It is handled automatically during estimation. Fixed common item calibration is a slight variation of this procedure that also places parameters on a common scale during the estimation routine. In this procedure, common item parameters on Form X are fixed to their estimated values on Form Y.

In separate calibration, parameters for each form are estimated separately and an additional step is needed to link estimates to a common scale. A consequence of setting the mean person ability to zero and standard deviation to one during separate estimation of Form X and Form Y parameters is that examinees taking Form X will have the same mean ability level as those taking Form Y even though the two groups may not be equivalent. That is, we end up with within group scales. To adjust the Form X parameters, we use the linear transformation $\theta_{Y*} = A\theta_X + B$ to place a Form X ability, θ_X, on the scale of Form Y. Similar transformations are applied to the item parameters. Discrimination is transformed by $a_{Y*} = a_X / A$ and difficulty is transformed by $b_{Y*} = Ab_X + B$ where the items parameters with an X subscript are parameters that belong to Form X. A and B are transformation coefficients derived from the common item parameters, and there are four popular methods for computing them (Hanson & Béguin, 2002).

MOOCs, on the other hand, enroll tens of thousands of students, which is more than enough to obtain accurate estimates with any IRT model. Large class sizes are one reason why IRT and MOOCs are the perfect marriage.

The mean/sigma (Loyd & Hoover, 1980) and mean/mean (Marco, 1977) methods are referred to as method of moments procedures because they use only item parameter descriptive statistics to compute the transformation coefficients. They are easy to implement and can be computed by hand. For example, mean/sigma transformation coefficients can be computed from the summary statistics in Table 2. The slope coefficient is computed from the common item estimates by dividing the standard deviation of Form Y item difficulty by the standard deviation of Form X item difficulty $A = \sigma(b_Y) / \sigma(b_X)$. The intercept coefficient is the mean item difficulty of Form Y subtracted by the rescaled Form X mean item difficulty, $B = \mu(b_Y) - A\mu(b_X)$. Using Table 2, these coefficients are $A = 0.51 / 0.58 = 0.88$ and $B = -0.02 - 0.88(-0.65) = 0.55$. The slope coefficient differs slightly from the value reported for the mean/sigma method in Table 2 because of rounding. The values in Table 2 are more accurate. The mean/sigma method gets its name because it uses the mean and standard deviation of item difficulty parameters. The mean/mean method, on the other hand, only uses the item discrimination and item difficulty means. It does not involve the computation of standard deviations. Specifically, the slope coefficient for the mean/mean method is $A = \mu(a_X) / \mu(a_Y)$. The intercept is computed in the same way as in the mean/sigma method. Using the values in table 2, the slope is $A = 1.03 / 1.22 = 0.84$ and the intercept is $B = -0.02 - 0.84(-0.65) = 0.53$. These values are slightly different from the tabled values due to rounding.

Table 2
Common Item Descriptive Statistics and Transformation Coefficients

	Form X Item Parameters			Form Y Item Parameters		
Statistic	*a*	*b*	*c*	*a*	*b*	*c*
Mean	1.03	-0.65	0.17	1.22	-0.02	0.17
S.D.	0.16	0.58	0.08	0.18	0.51	0.09

Method	*A*	*B*
Mean/sigma	0.89	0.55
Mean/mean	0.84	0.52
Haebara	0.87	0.53
Stocking-Lord	0.85	0.52

Note: transformation coefficients computed at full precision. Computing coefficients for the moment methods using descriptive statistics in the table above may differ slightly due to rounding.

Method of moments procedures are attractive because of their simplicity, but their main limitations are that they do not use all of the item characteristics and they can be affected by outliers. Alternatively, the Haebara (Haebara, 1980) and Stocking-Lord procedures (Stocking & Lord, 1983) are referred to as characteristic curve methods because they use item and test characteristic curves to obtain the transformation coefficients. Characteristic curve methods are computer intensive and require specialized computer software such as STUIRT (Kim & Kolen, 2004), the plink package in R (Weeks, 2011), and jMetrik (Meyer, 2013). Stocking-Lord and Haebara transformation coefficients are listed in Table 2. We used coefficients from the Stocking-Lord procedure to transform Form X parameters to the scale of Form Y (see Table 3). Parameters estimates in Table 3 are now on a common scale.

Table 3
Item Parameters after Linking with Coefficients from the Stocking-Lord Procedure

	Form X Item Parameters			Form Y Item Parameters		
Item	*a*	*b*	*c*	*a*	*b*	*c*
1	**1.38**	**1.00**	**0.11**	**1.31**	**1.09**	**0.10**
2	1.32	-0.66	0.24	1.25	0.35	0.22
3	1.03	-1.52	0.25	1.50	1.48	0.02
4	1.27	-1.92	0.24	1.44	-0.78	0.29
5	**1.12**	**-0.25**	**0.19**	**1.09**	**-0.30**	**0.23**
6	1.19	0.33	0.19	1.34	-0.30	0.23
7	1.22	1.49	0.02	0.96	0.15	0.25
8	1.36	0.66	0.07	1.22	-0.62	0.22
9	**1.06**	**-0.21**	**0.20**	**1.14**	**-0.01**	**0.26**
10	1.22	0.86	0.23	1.28	-0.44	0.12
11	1.14	0.31	0.30	1.28	-0.01	0.23
12	1.51	1.50	0.23	1.19	-0.86	0.14
13	**1.25**	**0.19**	**0.11**	**1.22**	**0.13**	**0.06**
14	1.38	0.32	0.13	1.26	0.07	0.21
15	1.14	-0.92	0.25	1.39	0.18	0.18
16	1.24	1.32	0.25	1.36	-0.23	0.10
17	**1.50**	**-0.49**	**0.33**	**1.53**	**-0.59**	**0.29**
18	1.52	1.21	0.16	1.06	0.37	0.06
19	1.03	1.29	0.22	0.90	0.46	0.17
20	1.11	0.24	0.04	1.14	1.46	0.12
21	**0.90**	**-0.55**	**0.19**	**0.95**	**-0.43**	**0.26**
22	1.09	-1.42	0.23	1.08	1.20	0.07
23	1.50	0.73	0.09	1.18	-1.00	0.30
24	1.22	-2.21	0.23	0.98	0.10	0.05
25	**1.13**	**-0.04**	**0.10**	**1.14**	**-0.07**	**0.11**
26	1.10	-0.54	0.23	1.10	0.50	0.10
27	1.15	0.88	0.22	1.04	0.61	0.26
28	1.17	0.17	0.17	1.10	0.15	0.14
29	**1.34**	**0.09**	**0.15**	**1.36**	**-0.02**	**0.10**
30	1.17	0.11	0.26	0.97	-1.07	0.24
Mean	1.23	0.07	0.19	1.19	0.05	0.17
S.D.	0.16	0.99	0.08	0.17	0.67	0.08

Note: Bold font indicates common items.

Despite the increasing impact of MOOCs on higher education, cheating poses a threat to online assessments in these courses.

Among the various methods for scale linking, the Stocking-Lord procedure works best when items are all of the same type (Baker & Al-Karni, 1991; Wells, Subkoviak, & Serlin, 2002), and the Haebara method works best in mixed format tests such as those that combine multiple-choice and short answer type items (Kim & Lee, 2006). Concurrent calibration and fixed common item procedures also work very well, particularly compared to the method of moments procedures. However, these two methods make it difficult to detect items that have an undue influence on linking process.

Score Equating with Item Response Theory

Testing programs report scores to examinees in a scaled score metric that is usually limited to positive whole numbers. For example, the GRE Verbal Reasoning scaled score consists of one point increments between 130 and 170. The purpose of scaled scores is to distinguish them from simple sum scores and have a metric that is independent of test forms. They are obtained by either transforming an examinee's IRT ability estimate or an examinee's sum score. In the former case, no further work is needed to produce comparable scaled scores; the linking process has already adjusted for difference among test forms and placed parameters on a common scale. IRT ability parameters are simply transformed to the scaled score and the work is done.

Recall that IRT ability parameters are invariant in different samples of items but observed scores are not. As such, if the scaled score scale is defined as a transformation of the observed score, an additional equating step is needed to adjust test forms for differences in difficulty. True score equating and observed score equating are two options in an IRT framework. True score equating is easier to implement as it involves test characteristic curves from two forms. As illustrated in Figure 4, Form X is easier than Form Y at low levels of ability, but at high levels of ability, the opposite is true. To adjust for these differences in test difficulty, we find the Form Y equivalent of a Form X score. As illustrated by arrows in Figure 4, the steps involve (a) choosing a Form X true score value (21 in Figure 4), (b) finding the Form X ability level that corresponds to that true score (0.61 in Figure 4), (c) computing the Form Y true score at the Form X ability level (22.3 in Figure 4). Thus, a Form X true score of 21 is equivalent to a rounded Form Y true score of 22.

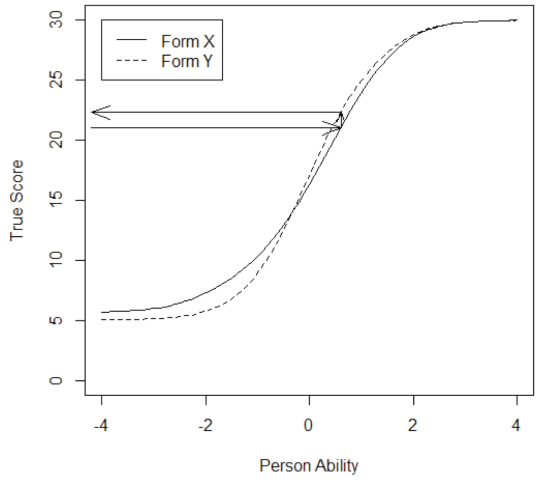


Figure 4. *An illustration of true score equating.*

Although true score equating is easy to illustrate, it actually requires computer intensive methods to implement. POLYEQUATE (Kolen & Cui, 2004), plink (Weeks, 2011), and jMetrik (Meyer, 2013) are three free programs that implement true score equating. Table 4 lists all of the equated true score values for Form X and Form Y. Scores from the two different test forms are now comparable. They have the same meaning and lead to fair and equitable decisions about student performance.

Discussion

Despite the increasing impact of MOOCs on higher education, cheating poses a threat to online assessments in these courses. Students may get illicit help via communication devices or even get access to answers before the assessment. Multiple test forms and extensive item pools can improve test security and increase fairness in online testing, but they leave open the possibility that test forms will differ in difficulty and give an advantage to students completing the easier form. Scale linking and score equating procedures must accompany the use of multiple test forms to ensure comparability among scores. Classical test theory methods commonly used in traditional course assessment can be extended to classical methods of score equating. However, these methods suffer from limitations such as population dependence. Large class sizes that are typical for MOOCs make a wide range of IRT models available for online assessment. IRT based scale linking and score equating overcome many of the problems with classical methods and make scale linking and score equating relatively easy to implement in practice.

Using a single counter–measure to combat cheating is like protecting your home from burglary by locking the doors and leaving the windows open.

Multiple test forms prevent unfair advantages due to prior knowledge of test items and the sharing of answer keys, but they do not prevent all forms of cheating. Indeed, using a single countermeasure to combat cheating is like protecting your home from burglary by locking the doors and leaving the windows open. An effective testing program makes use of multiple countermeasures to address all points of vulnerability. Multiple test forms should be combined

Table 4
True Score Equating Results

Form X True Score	Form X Theta	Form Y True Score Equivalent	Rounded Equivalent
0	-99.0	0.00	0
1	-99.0	0.89	1
2	-99.0	1.78	2
3	-99.0	2.67	3
4	-99.0	3.56	4
5	-99.0	4.45	4
6	-2.98	5.14	5
7	-2.15	5.58	6
8	-1.70	6.28	6
9	-1.36	7.26	7
10	-1.08	8.46	8
11	-0.86	9.79	10
12	-0.66	11.19	11
13	-0.49	12.60	13
14	-0.33	13.99	14
15	-0.18	15.35	15
16	-0.04	16.66	17
17	0.09	17.91	18
18	0.22	19.10	19
19	0.35	20.23	20
20	0.48	21.29	21
21	**0.61**	**22.30**	**22**
22	0.74	23.25	23
23	0.87	24.15	24
24	1.02	25.01	25
25	1.17	25.84	26
26	1.34	26.66	27
27	1.54	27.46	27
28	1.79	28.28	28
29	2.17	29.10	29
30	99.00	30.00	30

Note: Bold font indicates the true score equating relationship illustrated with arrows in Figure 4.

Accomplishing this criterion in practice will drive more institutions to offer course credit for MOOC completion and further expand the influence of these courses on higher education throughout the world.

with other counter measures such as proctored testing to combat cheating in a comprehensive way. Fair and equitable testing is achieved by minimizing all forms of cheating and ensuring the comparability of scores from different test forms. Accomplishing this criterion in practice will drive more institutions to offer course credit for MOOC completion and further expand the influence of these courses on higher education throughout the world.

Limitations

We simulated the data in this paper using a 3PL model. We obtained parameter estimates reported in the tables with ICL (Hanson, 2002) and conducted the linking and equating procedures in jMetrik (Meyer, 2013). We used simulated data to demonstrate IRT, scale linking, and score equating. As such, the data perfectly fit the 3PL model and are void of the usual noise of real test data. Our data also make it appear that equating does not change scores by much. However, this result is not always the case. Scores could change more substantially with real test data and greater difference in test forms. However, the only way to know the extent of the change in scores is to conduct the complete linking and equating process.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28,* 147-162.

Baylari, A., & Montazer, G. A. (2009). Design a personalized e-learning system based on item response theory and artificial neural network approach. *Expert Systems with Applications, 36*(4), 8013-8021.

Bond, T.G., & Fox, Ch.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Challis, D. (2005). Committing to quality learning through adaptive online assessment. *Assessment & Evaluation in Higher Education, 30*(5), 519-527.

Chen, C., Lee, H., & Chen, Y. (2005). Personalized e-learning system using item response theory. *Computers & Education, 44*(3), 237-255.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), *Educational measurement,* 4th Edition (pp. 355-386). Westport, CT: Praeger.

Cook, L. L., & Eignor, D. R. (1991). An NCME module on IRT Equating methods. *Educational Measurement: Issues and Practice, 10*(3), 191-199.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it.* Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J. (2003). *Detecting and preventing classroom cheating: Promoting integrity in assessment.* Thousand Oaks, CA: Corwin Press.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144–149.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhoff.

Hanson, B. A. (2002). *IRT command language* [computer software]. Retrieved from http://www.b-a-h.com/.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3-24.

Kim, S., & Kolen, M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models* [computer software]. Retrieved from http://www.education.uiowa.edu/centers/casma/computer-programs.aspx

Kim, S., & Lee, W. -C. (2006). An extension of four IRT linking methods for mixed format tests. *Journal of Educational Measurement, 43*, 53-76.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. *Applied Psychological Measurement, 11*(3), 263-277.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices.* New York, NY: Springer.

Kolen, M. H., & Cui, Z. (2004). POLYEQUATE [computer software]. Retrieved from http://www.education.uiowa.edu/centers/casma/computer-programs.aspx.

Kolowich, S. (2013, March). The professors who make the MOOCs. *Chronicle of Higher Education.* Retrieved March 20, 2013, from http://chronicle.com/article/The-Professors-Behind-the-MOOC/137905/#id=overview.

Lederman, D. (2013, February). Expanding pathways to MOOC credit. *Inside Higher Education*. Retrieved from http://www.insidehighered.com/news/2013/02/07/ace-deems-5-massive-open-courses-worthy-credit.

Lewin, T. (2013, February 20). Universities abroad join partnerships on the Web. *New York Times.* Retrieved from http://www.nytimes.com/2013/02/21/education/universities-abroad-join-mooc-course-projects.htm

Linacre, J. M. (2011). *Winsteps*® (Version 3.71.0) [computer software]. Beaverton, OR: Winsteps.com.

Liu, J., Sinharay, S., Holland, P. W., Curley, E., & Feigenbaum, M. (2011). Test score equating using a mini-version and a midi anchor: A case study using SAT® data. *Journal of Educational Measurement, 48*(4), 361-379.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179-193.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1–20. Retrieved from http://www.jstatsoft.org/v20/i09/

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139-160.

Meyer, J.P. (2013). *jMetrik version 3* [computer software]. Retrieved from www.ItemAnalysis.com.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57-75.

Muraki, E., & Bock, R. D. (1997). *PARSCALE 3: IRT based test scoring and item analysis for graded items and rating scales* [computer software]. Chicago, IL: Scientific Software International.

Reeves, T. C. (2000). Alternative assessment approaches for online learning environments in higher education. *Journal of Educational Computing Research, 23*(1), 101-111.

Rogers, C. F. (2006). Faculty perceptions about e-cheating during online testing. *Journal of Computing Sciences in Colleges, 22*(2), 206-212.

Rovai, A. P. (2000). Online and traditional assessments: What is the difference? *Internet and Higher Education, 3*(3), 141-151.

Rowe, N. C. (2004). Cheating in online student assessment: Beyond plagiarism. *Online Journal of Distance Learning Administration, 7*(2).

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63-84.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.

Thissen, D. (1991). *MULTILOG: Multiple category item analysis and test scoring using item response theory* [computer software]. Chicago, IL: Scientific Software International.

Thrift, N. (2013, February 13). To MOOC or not to MOOC. *Chronicle of Higher Education*. Retrieved from http://chronicle.com/blogs/worldwise/to-mooc-or-not-to-mooc/31721

Underwood, J., & Szabo, A. (2003). Academic offences and e-learning: Individual propensities in cheating. *British Journal of Educational Technology, 34*(4), 467-477.

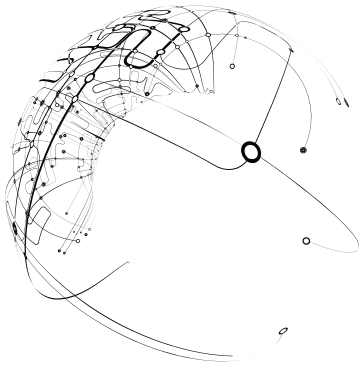
Wang, W. C., & Chen, C. T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models. *Educational and Psychological Measurement, 65*(3), 376-404.

Weeks, J. P. (2011). *Plink package* [computer software]. Retrieved from http://cran.r-project.org/web/packages/plink/index.html

Wainer, H., Dorans, N. J., Flaughter, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer,* 2nd Edition. Mahwah, NJ: Lawrence Erlbaum.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26,* 77-87.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items* [computer software]. Chicago, IL: Scientific Software International.



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.

Book Review

The One World Schoolhouse: Education Reimagined. Salman Khan. New York, NY and Boston, MA: Twelve Books, 2012. 259 pp. ISBN-13:978-1455508389. Hardcover, $26.99.

REVIEWED BY:
Catharine R. Stimpson, Ph.D.
New York University

"Bliss was it that dawn to be alive," wrote William Wordsworth, the canonical Romantic poet, "But to be young was very Heaven!" Born in 1770, he was remembering the joys of being an Englishman in France during the Revolutionary period.

Today, a tribe of exuberant, game-changing revolutionaries is storming, not the Bastille in Paris, but classrooms in America. Salman Khan is among the happiest and more attractive of these warriors. The One World Schoolhouse is his self-representation and a self-introduction to the world. It begins disarmingly, "My name is Sal Khan. I'm the founder and original faculty of the Khan Academy" (p. 1).

In part, the source of Khan's likeability is the amiable, plain populism of his ambition for his eponymous Academy, "To provide a free, world-class education for anyone, anywhere" (p. 221). Technology is the servant of this goal. If he succeeds, "tens of millions" of kids will gain access to education. The gap between rich and poor, between developed and developing societies, will vanish.

In part, the source of Khan's likeability is his temperament. Although his publisher claims that the destiny of his book is to be "one of the most influential… about education in our time," he is generous towards others, modest, and self-deprecating. He admits that his ambition might seem grandiose. Often wary of certainties, he resists believing that any one pedagogy---even his---will be best for one and all. He respects the complexity of the brain, the governor of learning. Unlike many reformers, he refuses to choose between the liberal arts and more utilitarian modes of education. Both have their virtues. He respects many teachers and shies away from bashing their unions. Although his manifesto is silent about the grand American pragmatic tradition, he often seems more pragmatist than revolutionary. "My personal philosophy," he writes, "is to do what makes sense and not try to confirm a dogmatic bias with pseudoscience" (p. 131). Winningly, his pragmatism is joyous rather than cramped, for he celebrates the wonder, the excitement, the "magic" of learning.

The One World Schoolhouse seeks, even strains, to be earnestly conversational in tone. The book nevertheless echoes three mythic narratives that provide a ground bass in American culture. They resound beneath the four-part formal structure of the book: how he learned to teach; how broken our current educational model is; how he brought his ideas to "the real world"; and how his breathless vision of the future, that one world schoolhouse, might operate.

The presence of these mythic narratives is still another source of Khan's appeal. For his story fits with older, familiar tales that we hope might be true. One is the myth of origins, in which we learn about the beginnings of a hero who becomes capable of legendary deeds. Two others are less universal, more American, and echo each other. The first is that of the young men whom Horatio Alger (1832-1899) imagined so prolifically. Full of pluck, bestowed with some luck, they journey from rags to riches. Since the late 19th-century, such exemplary figures have become far more multi-cultural, including African Americans, immigrants, and even some women.

The second narrative focuses on the tinkerer, usually a man, fooling around in his barnyard or garage or kitchen, often in isolation, the beneficiary of "serendipity and intuition" (p. 33). If his experiments lead to inventions, and if he is entrepreneurial, he starts up a little company, and if he also has pluck and a dose of luck, he will build a Ford Motor Company, or a Hewlett Packard, or a Khan Academy. As of May 2013, the Khan Academy claims to have delivered over 250 million lessons - in English and a variety of other languages.

Born in Louisiana, Khan is the son of immigrants, his father a pediatrician from Bangladesh. He is reticent about his childhood, but one sentence points to family difficulties. Both he and his wife Umaima, he writes, "come from single-mother households whose earnings were slightly above the poverty line in a good year…" (p. 154). He goes to MIT, becomes a hedge fund analyst, and marries a doctor. At his wedding in 2004, he meets a young female cousin, Nadia, a bright girl who has done badly on a 6th grade placement examination in math. He volunteers to tutor her. After improvised math lessons delivered through computer, pen tablets, and long distance phone calls, Nadia successfully passes the test.

His lessons last for only ten minutes, because that is the time limit for a YouTube posting, but lo and behold, ten minutes is the length of his students' attention spans.

This charming act of benevolence is the start of the Khan Academy. Through word of mouth among family and friends, the number of his tutees grows. As it does, the teacher changes. Psychologically, he realizes he has a vocation, a passion to pursue. Pedagogically, he writes new software, which improves his questions and his ability to follow his tutees' answers. To manage the scale of his still pro bono enterprise, he posts his lessons on YouTube. He has no formal training as a teacher, but he is smart, caring,

and not afraid to fail. His lessons last for only ten minutes, because that is the time limit for a YouTube posting, but lo and behold, ten minutes is the length of his students' attention spans.

In 2009, with the encouragement of his wife and friends, he quits his secure job, and opens the Khan Academy, located in a closet in his home, now in California. At first only he is "faculty, engineering team, support staff, and administration" (p. 6), but he dares to dream of an educational transformation. Not only does he expand his curriculum to include basic arithmetic, calculus, physics, finance, biology, chemistry, and history. Not only does he attract millions of students to his lessons. Not only does he have the good sense to test out his methods in on-site programs. Not only does he build an organization. He attracts powerful and affluent supporters. One of them goes to the Aspen Ideas Festival and hears Bill Gates say that he is a fan of the Khan Academy, that he uses its videos for "his own learning and for his kids" (p. 158). Since the end of the 19th-century in the United States, foundations have provided much of the financial muscle for educational reform. As the Gates Foundation, Google, and other philanthropies offer their support, the Khan Academy shows their continued power.

Like most revolutionaries, Khan must dismiss the past in order to legitimate his brave new world that will replace it.

Although the Khan Academy depends on the charisma of its founder, it builds on three clusters of ideas, none original but articulated with buoyant, even breezy, enthusiasm. The first is a history of education, which blames "The Prussians" of the 18th century and their American acolytes for designing a rigid system that by mandate locks children into a classroom and then promotes them in lock-step from grade to grade. The "Prussian" legacy is a dangerously obsolete machine, incapable of stimulating curiosity and life-long learning, and carrying "… such a weight of orthodoxy and rust as to stifle the sincere creative efforts of even the best-meaning teachers and administrators" (p. 80). Colleges and universities, devoted to the "broadcast lecture" are equally deadening. I looked in vain for the names of such influential reformers as John Dewey (1859-1962) or Maria Montessori (1870-1952) or Jerome Bruner (1915---), but like most revolutionaries, Khan must dismiss the past in order to legitimate his brave new world that will replace it.

Far more appealing is Khan's Romantic picture of children. They are born intelligent, curious, with an active and natural love of learning. They should be like Alices in a benign wonderland. Though their schools balkanize knowledge, they delight in making connections. The more they learn, the more their brains, like those of adults, flourish---according to the contemporary neuroscience Khan uses. How, then, do they best learn? Khan's footnotes are sparse, but he does credit the Winnetka Plan of the 1920s and the psychologists Benjamin Bloom and James Block of the post-WWII period for the theory and practice of "mastery learning." No matter how long it might take, students should "adequately comprehend a given concept before being expected to understand a more advanced one" (p. 37). If they get stuck, they should struggle and push and prod themselves at their own pace until they get unstuck. Khan is fond of the homely metaphor of the "Swiss cheese brain." If we have mastered one part of a concept but not another, we have debilitating gaps and holes in our learning.

Holding the third cluster of ideas together is Khan's vision of the "One World Schoolhouse," a loving globalization of the one room schoolhouse of yore, with children of several ages sitting together on their benches, helping each other under the guidance of one teacher, a stripped-down community of learning. Khan rattles off suggestions for radical change in many current practices---such as tracking, homework, grading, testing, and the keeping of transcripts. But obviously, technology is the Driver, the Big Cheese, of revolution.

Technology enables "differentiated," or individualized, learning for students, each of whom has a "feedback dashboard" that shows in real time a leaping or crawling toward mastery. Because of technology, Khan can picture a large, cheery classroom with a team of teachers and students of various ages engaged in projects, including the arts. Khan is far more vocal about the dangers of age segregation within schools than neighborhood segregation among schools. However, because of technology, education can become more affordable, giving poorer kids the same advantages that richer kids now have. Because of technology, the classroom can be both local and global. "Imagine," Khan enthuses, assigning the One World Schoolhouse the ability to transcend national rivalries, "… students in Tehran tutoring students in Tel Aviv or students in Islamabad learning from a professor in New Delhi" (p. 252). Presumably, they would find a common language in which to communicate.

Refreshingly, Khan is suspicious of the conviction that technology alone is the Super Fix of educational disrepair. Technology, he insists, enhances rather than dominates learning.

Refreshingly, Khan is suspicious of the conviction that technology alone is the Super Fix of educational disrepair. Technology, he insists, enhances rather than dominates learning. Enlightened educators integrate it in "meaningful and imaginative ways" (p.122). So arguing, Khan preserves an honored role for teachers. They coach; they mentor; they inspire; they provide perspective. Both students and teachers benefit from "face time." (Reading Khan I must face up to some of the more egregious rhetoric common to current educational reform.) "Face time" happens after students have

used the Khan Academy introduction to mastery learning, and when it does, sweetness and strength flow. For "face time shared by teachers and students is one of the things that humanizes the classroom experience, that lets both teachers and students shine in their uniqueness" (p. 35).

Likeable though Khan can be, *The One World Schoolhouse* is also irritating. Written for a general audience, it indulges in slapdash generalizations about psychology and history. For example, the remarks about the early university, which did train poorer boys for good careers, are silly. "Early universities pursued esoteric topics for a handful of privileged people who'd done their early learning at home; most of these students were wealthy or connected enough that 'work' was almost a dirty word" (p. 75). Perhaps not surprisingly, he ignores the contemporary university as a source of the concepts that children should learn. Indeed, a disturbing feature of much writing about radical change in education is an apparent indifference to the wellsprings of discoveries, new questions, and fresh ideas, and primary among these wellsprings is the university. It is all very well to praise student-centered learning. It is all very well to deploy technology-enhanced methods and metrics in pursuit of it. It is all very well for young men and women who already possess intellectual and social capital to scoff about going to college. However, students need to be learning something. What is the content of Khan's thousands of videos? If the subject at hand is the French Revolution, and if the concept at hand is "revolution" or "historical change," the research universities provide our agile, informed experts for both.

It is all very well to praise student-centered learning. It is all very well to deploy technology-enhanced methods and metrics in pursuit of it. It is all very well for young men and women who already possess intellectual and social capital to scoff about going to college. However, students need to be learning something.

Even more irritating is the comparative narrowness of Khan's chosen focus on students' lives. The social and economic facts about the context of these lives are stubborn things. Khan is hardly socially obtuse. He mentions global "poverty, hopelessness, and unrest" (p. 221). His paragraphs about the Indian subcontinent are alert to child malnutrition, a weak infrastructure, and administrative laxness and corruption. His sincerity about "making a difference" for all children is palpable. He wants them to be kind, good, thriving, well-educated global citizens.

However, his compelling interest is in the schoolroom and not in the home, or neighborhood, or church, or school board that surrounds it. Crucially, a child can be passive at her desk because she is being brutally sexually abused at home, not because she is a victim of rote learning. As a result of his focus, Khan's descriptions and prescriptions can lack the force of other important books about reform that have a wider-angled lens. I think, for example, of Patricia Albjerg Graham's *S.O.S. Sustain Our Schools,* published in 1992 but still relevant. It is blunt about the need for change. "Today the practices of the past are inadequate for the present and irrelevant for the future" (p. 17). Yet, she puts education and its remedies into a social landscape, "a deterioration in the lives of many children, exemplified by increased poverty, unstable families, and reckless consumerism" (p. 17).

My criticism will seem like carping if the Khan Academy leads diverse people of all ages, outside and inside of the formal classroom, to an education that is cognitively wide and deep; imaginatively creative and engaging; and morally resonant. The benign glories of the global schoolhouse itself cannot be assessed. They are too far in the future, too visionary, too sketchy, too blue sky. However, Khan seems eager to have his pedagogy robustly examined, even if his methods of assessment are but skimpily and loosely mentioned. Part of the problem of the assessment of Khan is similar to the problem of the assessment of any pedagogy that a provider delivers from his or her self-constructed platform to anyone who wants to download it. How does any objective observer insert him or herself into the process and measure what is really going on, using transparent criteria? Discern the efficacy of the lessons, the videos and problem sets and feedback mechanisms? Another part of the problem is that Khan promises, not only that people will learn, that they will master a concept, but that they will feel better about learning, happier and more self-confident. How does any objective observer measure individual character growth?

Khan seems to assume blithely, but not stupidly, that the popularity of his lessons, in several languages, is one important proof that people benefit from them cognitively and psychologically. He has tributes from his students. To suggest an analogy: if lots of people eat Cheerios, and some people write to the manufacturer and say that their children adore them, Cheerios must be nutritionally good for you. Khan does explicitly argue that his lessons have gotten extensive field-testing in the few years that the Khan Academy has existed. Through guesswork and experimentation, again mentioned rather than analyzed, he has come to believe that students have mastered a concept when they can "correctly solve ten consecutive problems on a given subject" (p. 138). He also notes some programs that have had a group using Khan techniques and a control group that did not. The test scores in the Khan group, he reports, increased significantly.

Measuring the Khan Academy critically will be more common if and when more existing classrooms collaborate with it in a systematic way. Assessors can get inside the process more easily. A straw in the winds of revolutionary change is in Idaho. In late February 2013, it announced a

pilot program involving 47 schools—charter, public, and private---that are to use Khan Academy materials. The J.A. and Kathryn Albertson Foundation is to give $1.5 million to these schools, in part for assessment. Moreover, the Foundation is donating money to the Northwest Nazarene University to support this activity through its Center for Innovation and Learning, previously unknown to me. Located in Nampa, Idaho, Northwest Nazarene is a private, liberal arts college, associated with the Church of the Nazarene, which also offers a handful of master's programs and one doctorate, an Ed.D.

However, his compelling interest is in the schoolroom and not in the home, or neighborhood, or church, or school board that surrounds it.

We shall see what we shall see, and we had better look. Meanwhile, the Khan Academy charges on, and Sal Khan charismatically spreads his gospel from the multiple platforms of contemporary communications. Instructively, the subtitle of his book is "Education Reimagined," far more Romantic and less technocratic than such favored, but less glamorous and dramatic slogans of great change as "reengineering." Because this is contemporary education, the often cheesy amalgamation of commerce and branding is never far from seductive promises of revolutionary academic change. The Khan Academy has a website, of course, and on that site is an official on-line store. One can purchase a Khan Academy onesie for $19.50, a Union Square Laptop Messenger bag for $75.50. Learn more, buy more. Examining these linked imperatives of cultivating the mind and spending money, a primary feature of our "revolutionary" educational moment, calls for the moral and political talents of all of us.

References

Graham, P.A. (1992). *S.O.S. Sustain Our Schools*. New York, NY: Hill and Wang.

Book Review

Measuring College Learning Responsibly.
Richard J. Shavelson. Stanford, CA:
Stanford University Press, 2009. 207 pp.
ISBN-13:978-1455508389. Hardcover, $26.99.

REVIEWED BY:
Jeff Kosovich, M.A.; Rory Lazowski, Ed.S.;
Oksana Naumenko, M.A.; & Donna L. Sundre, Ed.D.
James Madison University

Editor's Note: The following review has been co-authored by three graduate students and their instructor. Please note the introduction and conclusion are from the point of view of the faculty member and the body of the review is co-authored by the graduate students.

Richard Shavelson's book, *Measuring College Student Learning Responsibly*, was an answer to a real need for my graduate Psychology 812 course, "Assessment Methods and Instrument Design," a core requirement for Quantitative Psychology concentration Master's and Assessment and Measurement PhD students. A weekly course feature is student written reflections on each assigned reading. The book promised to discuss assessment, accountability and accreditation in the United States and to provide an international perspective. Given Rich Shavelson's prominence as a researcher and instrument developer, the book beckoned. This book promised to fuel our weekly seminar conversations and to provide just the kind of heat and controversy to inspire deep learning and engagement. When offered the opportunity to review this book, the perspectives of my fall 2012 students seemed the ideal ingredient; three of the best students from that cohort were recruited. Throughout the course, their unique perspectives were inspiring and remind us that we are all students together. Enjoy, as I did, the thoughts, reflections, and, yes the rants, of these students as they team to review each chapter.

Assessment and Accountability Policy Context

In the opening chapter, Shavelson wastes little time delving into the impact of the Spellings' Commission (U.S. Department of Education, 2006) recommendations and their impact on higher education assessment activities. One of the most salient issues that Shavelson examined is the continuing "tug of war" among the various cultures and stakeholders involved in assessment: the academic culture, clients (parents, students, government agencies, businesses), and policy-makers. Shavelson notes that the Commission's report took a client-centered stance, noting that universities should be transparent about student learning, the inherent value-added of attending a university, and the outcomes associated with a costly education. Simply put: accountability. It appears as though those in the policy-maker community share this same perspective. This stance is in contrast to the academic culture, which largely views assessment as leading toward instructional improvement in teaching and learning. This chapter outlined many of the recommendations in the Spellings report and how these recommendations were met by each culture.

Another central concept is the need for institutions to focus on sound assessment methodology. A recurring, albeit previously unheard of theme Shavelson promotes is that institutions may not be ultimately judged on assessment outcomes, but instead by assessment program quality. This is important because many institutions may fear that unsatisfactory results will lead to undesirable repercussions. Keeping quality practice at the forefront can assuage fears and influence more positive engagement in assessment practices. This chapter provides a focused and cogent treatment to some of the most persistent and pervasive issues in higher education assessment, issues that are more fully developed and explained in subsequent chapters.

Framework for Assessing Student Learning

Chapter 2 provides a host of information to consider for best practices in assessment, from how students should be assessed to the range of knowledge, skills, and abilities to be incorporated into an assessment plan. It further guides the reader through proposed measurement and student learning best practices for both accountability and improvement purposes. This treatment provides an accessible framework for designing an assessment plan and ways to improve existing plans.

A recurring, albeit previously unheard of theme Shavelson promotes is that institutions may not be ultimately judged on assessment outcomes, but instead by assessment program quality.

Shavelson outlines several key considerations to student learning assessment. He introduces an essential and critical distinction between learning and achievement: learning requires the observation and measurement of cognitive growth over time, while achievement provides only a single assessment of cognitive ability. Another important assessment best practice was offered through the distinction of the definitions and value of direct vs. indirect measures of student learning. Moreover, Shavelson recommends assessment designs that allow for both comparability across institutions and diagnosis within. This chapter describes a learning outcomes framework and advises us not only on what should be assessed (including soft skills), but also on how to assess efficiently.

Brief History of Student Learning Assessment

In Chapter 3, Shavelson describes historical roots of assessment and notes significant trends for student learning

outcomes. Of particular note, Shavelson provides a very useful discussion concerning the transition from internal institutional initiation of assessment to more external demands. This transition is coupled with a review of the paradigm shifts that accompanied student learning assessment purposes. Shavelson provides an excellent summary of landmark tests from the past century (e.g., the Pennsylvania Study, 1928-32; GRE) and notes that past perspectives of learning, thinking, and the purpose of higher education are still present in the way we measure learning today. However, Shavelson contends that standardized tests are incapable of providing granular data that can impact instruction and actual learning. He believes internal instruments are necessary to supplement the blunt standardized tools used for accountability purposes. This represents a challenge that few assessment practitioners will be able to address efficiently or effectively. Curiously, Shavelson neglects discussion on the topic of performance assessment until late in the chapter despite its long history (Johnson, Penny, & Gordon, 2008). However, he notes an important paradigm shift in what is considered a critical outcome of higher education: fact-based and procedural knowledge was of past value; broad skills and abilities such as critical thinking and verbal reasoning currently dominate today's standardized testing contexts. It is here that he introduces the College Learning Assessment (CLA). The CLA represents one example of large scale assessment that avoids the multiple choice format; however, the level of consideration it receives in this chapter and the rest of the book cumulatively becomes more of a very thinly-veiled attempt at product placement than an objective treatment of student learning assessment.

The Collegiate Learning Assessment

Chapter 4 focuses entirely on the development, philosophy, and strengths of the CLA, often citing promising results that require further investigation. Shavelson wooingly showcases the "prominence" of the CLA, framing it within the context of reliability and validity. First, although some of the presented reliability values are acceptable, those at the lower end of the range are quite frightening for a potentially high-stakes test. Given that test results may be used to provide information about school accreditation, funding, and diagnostic decision-making, it is necessary to ensure that scores are of high quality (i.e., good reliability and validity). Second, the use of "holistic" scores reported by the CLA seems to ignore the benefits and recommendations of the field of psychometrics, as well as the earlier plea for more granular, diagnostic data to drive learning improvement. Even if the test operates on the assumption that the test is 'greater than the sum of its parts,' the parts are not to be discarded. What does a total score of 1150 mean? This is the heart of validity. Third, evidence of high student motivation to perform well on this instrument appears to be assumed. Shavelson defends the CLA by arguing that test motivation is a concern with any standardized test; however, poor motivation, particularly on arduous performance tasks, brings the validity of test scores into question.

Finally, the CLA is said to be useable as an index of value-added, which is problematic. In terms of psychometrics, difference scores (i.e., the differences between test scores at two time points) tend to be even less reliable than test scores. This lack of reliability is problematic because the desire to compare schools on value-added is a major driving force in higher education accountability. If the test scores are inconsistent—and the value added scores are even less consistent—should we tempt policy makers to use these scores for accountability purposes? The problem with

Pitching the CLA as a measure of value added is essentially implying that the CLA measures the whole of student learning in some meaningful way. Shavelson repeatedly acknowledges that the CLA requires better validity evidence, yet he presents the CLA as a panacea to the woes of higher education assessment.

the value-added scores is further compounded by the way in which value-added is defined. Pitching the CLA as a measure of value added is essentially implying that the CLA measures the whole of student learning in some meaningful way. Shavelson repeatedly acknowledges that the CLA requires better validity evidence, yet he presents the CLA as a panacea to the woes of higher education assessment. Given Shavelson's expertise in psychometric matters (e.g., Shavelson & Webb, 1991), it is disappointing to see gross misrepresentation of the CLA's quality.

Exemplary Campus Learning Assessment Programs

In Chapter 5, Shavelson discusses the internal (assessment for improvement) and external (accountability for comparison) foci of assessment, displaying the diversity of assessment practice through the discussion of two named and four unnamed universities (one of which was clearly ours) that serve as exemplars of particular assessment practices. These profiles were useful in identifying distinctive characteristics and attributes among existing assessment programs and provide a convenient summary table (pp. 81-82). After describing each model in detail Shavelson came to several conclusions: 1) it is campus leadership that assures and inspires quality assessment programs, not accreditation agencies; 2) although all programs were outcomes-based, they significantly differ in the focus of their programs; 3) faculty engagement is critical; 4) explicit, measureable learning outcomes are key to appropriate use of data; 5) successful assessment programs require champions from diverse perspectives; 6) data must provide relevant information to guide faculty; 7)

incentives are required for policy, hiring, and rewarding assessment practice; and 8) practice must be balanced and sustainable to thwart morale problems. The value of the chapter is in the elucidation of several characteristics that differ broadly across existing model programs. This chapter was a highlight of the book and should have great utility for practitioners and those planning assessment centers.

The Centrality of Information in the Demand for Accountability

In his discussion of the centrality of information in Chapter 6, Shavelson briefly lays out the nature of higher-education accountability. Overall he does an excellent job introducing readers to the purpose of accreditation as well as the role of accrediting bodies. More importantly, Shavelson gives due treatment to the summative versus formative debate that has haunted assessment since the very beginning. The business-driven philosophies that have contributed to continued socio-political conflict within assessment are also examined. The bulk of the conflict appears to stem from the different types of information demands from various stakeholders. A particularly interesting contrast is made between politicians and consumers (e.g., students, parents). Shavelson also gives a nod to the Voluntary System of Accountability (VSA). His discussion of the VSA is one of the weaker points of the chapter because it is only portrayed in a positive light. Readers wooed by the VSA coverage should consider its success and shortcomings more deeply before endorsing it outright. For example, the participation rate in the VSA has plummeted since its inception. Ironically, much of the disillusionment with the VSA stems from its very reliance on standardized tests (Jankowski et al., 2012) that Shavelson himself earlier laments (while, even more ironically, steadfastly endorsing the CLA). In general, Chapter 6 serves as a primer for the later portions of the book dedicated to the topic of accountability.

Accountability: A Delicate Instrument

This chapter addresses complications that arise in higher education accountability efforts. At face value, accountability may appear reasonable and practical; however, these complications can lead to unintended consequences if applied inappropriately. Shavelson discusses six complications that underscore the notion that accountability is both a powerful tool and a delicate instrument. The complications include: 1) accounting for assessment outcomes vs. processes; 2) consideration of what processes and outcomes to measure; 3) problems associated with the use of sanctions (and/or incentives) as a vehicle for change or improvement; 4) the functions of formative vs. summative assessment purposes to inform multiple audiences; 5) appropriate and ethical interpretation of results; and 6) the balance between complying with external expectations and maintaining flexibility and innovation in assessment efforts. These complications highlight more global issues related to accountability that are important considerations. Although we perseverate on development of more specific measurement and instrument design (with good reason), this chapter helps to identify the "forest" issues that may hide the accountability system "trees." The reader can take a step back and reflect on how delicate these considerations may be and how important it is to think intelligently about these more overarching accountability issues. This chapter is highly recommended for policy-makers.

State Higher-Education Accountability and Learning Assessment

Similar to the previous chapter, chapter eight provides another global view of the issues in higher-education learning assessment and accountability with a focus on states' influences. Shavelson delves further into accountability policy by examining how learning is assessed in US higher education. One of the central points of the chapter is the catch-22 facing higher education institutions.

The mad dash by various organizations to compare and rate the effectiveness of institutions often leaves stakeholders with piles of uninterpretable numbers.

The mad dash by various organizations to compare and rate the effectiveness of institutions often leaves stakeholders with piles of uninterpretable numbers. For example, the sheer volume of learning indicators can result in different institutions measuring different constructs with different instruments, making any meaningful comparison impossible. To make matters worse, the indicators reported (e.g., graduation/retention, enrollment, tuition) essentially define what is and is not important in assessment in the eyes of some stakeholders. This cycle of confusion and disorganization leads to wave after wave of expensive, loosely-aimed assessment initiatives. These initiatives provide minimally-useful information while simultaneously contributing heavily to institution-level decisions. The focus on measures needs to be shifted from indirect measures, such as graduation rates and retention, to more direct measures of learning. Overall the chapter provides an even-handed treatment of the issues. Unfortunately, the topic of direct measures opens up Shavelson's obligatory CLA pitch for the chapter.

Higher Education Accountability Outside the United States

This chapter outlines international assessment of student learning. Countries other than the U.S. that are covered in this chapter approach accountability in qualitatively different ways by employing a four-stage quality assurance process rather than isolated assessment programs. Instead of looking at several different input and

output indicators, the countries that address accountability focus more on "quality assurance processes." Shavelson describes these as more focused on the processes that ensure quality education. This diagnostic, monitoring approach has many advantages over the U.S. tradition including continuous access to evidence-based practices, by advancing and reinforcing continuous learning improvement. Additionally, Europe and Australia do not appear to wage wars on colleges and universities within this system, though their faculty may disagree. Shavelson is correct in pointing out that these quality assurance practices seem sustainable and should be noted by American policy-makers. He explains why the academic audit stage is the one stage from which the U.S. can learn. In the context of other nations' federal burden of funding higher education outside of the U.S., it appears justified that universities conform to external statures. Shavelson highlights a stark contrast between other nations' explicit responsibility in funding higher education and the invasive nature of the U.S. government policies, which are responsible for only small portions of higher education funding. Overall, this chapter represents an excellent example of "compare and contrast" between higher education learning measurement across nations from which policy makers can and should learn.

Learning Assessment and Accountability for American Higher Education

This final chapter nicely ties together information discussed in previous chapters to form a vision of assessment in higher education. In doing so, Shavelson integrates the existing body of knowledge presented in the book and attempts to conceive a better system for United States learning assessment and accountability. He effectively addresses many of the extant tensions, with potential solutions to these tensions, both in terms of substance (what should be measured) and structure (how to measure it). Whether or not it can be achieved is another question. However, credit is due for putting forth pragmatic strategies for resolving the reconciliation of assessment and accountability issues.

It would have been helpful to provide alternative assessment options for those who are not interested in using or cannot afford the CLA.

In yet another attempt to advertise, Shavelson recommends the CLA to achieve this vision of assessment in higher education. It would have been helpful to provide alternative assessment options for those who are not interested in using or cannot afford the CLA. Nevertheless, there are some clear guidelines set forth regarding what should be measured and how to measure it. This can be particularly valuable to assist institutions struggling to improve in their assessment and accountability efforts.

Summary

You may agree or disagree with the views provided by these graduate students; however, it cannot be ignored that Shavelson's book provided a powerful accelerant to provide not just a fire–perhaps a bonfire–of graduate level discussion. On several occasions we basked in the heat and glow of these conversations discussing all of the topics so well covered. These are important topics for all interested in assessment and accountability; we invite you to join the conversation and make positive change. For those less desirous of potential conflict, there are at least two other viable choices for textbooks or learning more about higher education assessment practice. Linda Suskie's (2009) 2nd edition, fully lives up to its title, *Assessing Student Learning: A Common Sense Guide*, and provides what many would consider a desktop reference. Another excellent choice would be Astin and antonio's (2012) 2nd edition *Assessment for Excellence*. For a recent review of the latter by Linda Sax (2012), please look no further than *Research & Practice in Assessment.* Both books are excellent new additions to our assessment bookshelves.

References

Astin, A.W., & antonio, a. (2012). *Assessment for excellence: The philosophy and practice of assess ment and evaluation in higher education* (2nd Edition). Lanham, MD: Rowman & Littlefield and the American Council on Education.

Jankowski, N. A., Ikenberry, S. O., Kinzie, J., Kuh, G. D., Shenoy, G. F., & Baker, G. R. (2012). Transparency & accountability: An evaluation of the VSA College Portrait Pilot. (A Special Report from NILOA for the Voluntary System of Accountability). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. Retrieved from http://www.learningoutcomeassessment.org/documents/VSA_000.pdf

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks.* New York, NY: Guilford Press.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* New York, NY: Sage Publications, Incorporated.

Suskie, L. (2009). *Assessing student learning: A common sense guide.* (2nd Edition). San Francisco, CA: Jossey Bass.

U. S. Department of Education. (2006). *A test of leader ship: Charting the future of U.S. higher education.* Washington, DC: Author.

Book Review

We're Losing Our Minds: Rethinking American Higher Education.
Richard P. Keeling and Richard H. Hersh.
New York, NY: Palgrave Macmillan, 2011. 205 pp.
ISBN-13:978-0-230-33983-5 Paperback, $24.00.

REVIEWED BY:
Katie Busby, Ph.D.
Tulane University

My parents were not "helicopter parents" and when I went off to college they only supplied me with basic boundaries and an expectation that I would do my best and make good choices; the rest was up to me. When I returned home for visits my father and I would talk about my classes and extracurricular activities and during one of these conversations I casually asked him what he thought I should be learning in college. I expected his answer would include lessons from history or great works of literature, but he surprised me with a succinct yet complex response when he replied, "College is where you learn how to think." My father was a businessman and did not use the jargon of higher education. He would not have used phrases such as "critical thinking" or "higher learning" but as a businessman, a community volunteer, and a civic leader he knew the importance of these skills and in his own way he encouraged me to use my time in college to develop them. As my father impressed the importance of learning how to think upon me, Richard Keeling and Richard Hersh impress the importance of higher learning upon the readers of their book *We're Losing Our Minds: Rethinking American Higher Education.*

They freely acknowledge the challenges and shortcomings of colleges and universities, but do not dwell on these deficiencies. The authors provide concrete suggestions for improving the college experience for students and are optimistic and confident about these possibilities.

Throughout this book the authors cut to the chase, offering a candid and sometimes painful assessment of the state of higher education in the United States. However, their work is not limited to a litany of what is wrong with colleges and universities today. They freely acknowledge the challenges and shortcomings of colleges and universities, but do not dwell on these deficiencies. The authors provide concrete suggestions for improving the college experience for students and are optimistic and confident about these possibilities. Their solutions are anchored by the idea of a college environment that embraces a holistic approach to student learning and advances higher learning. This environment includes a culture of strong learning and teaching where students are apprentices and a comprehensive, intentional, integrated, and rigorous program of curricular and co-curricular experiences supports student learning and development.

Keeling and Hersh define higher learning as the "learning that prepares graduates to meet and excel at the challenges of life, work, and citizenship" (p. 1). Higher learning does not occur after a particular course has been completed or when a requisite number of service learning activities have been performed. Higher learning develops gradually over time, includes thoughtful reflection on curricular and co-curricular experiences, and requires the student to engage in his or her environment to develop a deeper understanding of self and make meaning of the world in which he or she lives. Higher learning is a distinguishing characteristic of higher education and more higher learning is needed.

Higher learning develops gradually over time, includes thoughtful reflection on curricular and co-curricular experiences, and requires the student to engage in his or her environment to develop a deeper understanding of self and make meaning of world in which he or she lives.

In the first chapter, Keeling and Hersh lament that learning no longer seems to be a priority on many college campuses and appeal for learning to be returned to that top position. The authors endorse the perspective that college prepares one for life and citizenship as they caution against a narrow characterization of higher education as simply a pathway to employment or the means to advance from an existing job to a better one. This characterization, coupled with the concerns over lack of job prospects for graduates, the rising cost of tuition, and the amount of debt accrued by students is commonplace in news reports and even finds its way into conversations taking place among students, their families, and other stakeholders. Keeling and Hersh emphasize that college costs are not the problem, but the lack of value students receive in return is. To resolve these concerns, the authors call for radical changes, not incremental ones, and readers can easily agree with the declaration. However, this suggestion can be perilous for today's academic leaders as evidenced by the resignation and subsequent reinstatement of University of Virginia's president in June 2012, where differences in opinion of how change should be managed and how quickly it should occur were at the heart of the matter (Hebel, Stripling, & Wilson, 2012).

In chapter two, Keeling and Hersh describe the arguably perverted nature of college metrics and rankings. They do not sugarcoat their feelings about glossy brochures and recruitment rhetoric and wonder why the focus is placed on satisfaction and services and not on the teaching and learning mission of colleges and universities. Directors of assessment,

institutional research and/or institutional effectiveness know the institutional counts and satisfaction ratings that abound in university publications, on websites and in various rankings are easier to collect, report, and compare than the results from direct assessments of student learning.

Keeling and Hersh focus on developmental learning in chapter three asserting that learning and development are inextricably intertwined and that learning cannot be discussed without also discussing student development. Unfortunately, the authors find that developmental learning is largely absent in colleges and universities because the integrative and purposeful collection of curricular and co-curricular programs that stimulate such learning are not offered in a holistic manner that supports higher learning. This shortcoming is compounded by the challenges of alcohol misuse/abuse, drugs, the prevalence of psychological and behavioral disorders, and risky behaviors that are present on college campuses and interfere with student development and learning. The authors describe student development through narrative examples which some readers may appreciate. However, others may dislike this approach and feel that Keeling and Hersh slight the existing student development research. Renowned theorists such as Perry, Chickering, Sanford, or Kohlberg, whose theories describe the student development taking place in the authors' examples are not explicitly mentioned. Instead the authors give only a brief nod to the "extensive scholarly literature" and a reference that includes a few prominent works.

The authors emphasize that higher learning requires knowledge acquisition as well as opportunities to apply, consider, challenge and make meaning of that knowledge in the context of the broader world. They believe this can happen best by immersing students in an environment that fosters both acquisition and application and allows students to develop into unique individuals who make meaning for themselves, and that faculty are best-suited to foster this environment. Keeling and Hersh acknowledge that this will take more time and effort, but passionately believe this investment will yield incredible dividends – for the student, for the faculty member, for higher education, and for society. The authors make a valid point, but it can take a long time for the dividends to be paid and maintaining the necessary level of investment can be difficult for overworked faculty and student affairs professionals who are expected to demonstrate results quickly.

Richard Keeling is a medical doctor so it is not surprising that chapter four focuses on the neuroscience of learning. The systemic nature of body, brain, mind, and learning are outlined clearly in a way that avoids the use of detailed medical jargon, but the key theme of learning as a physical process is slightly oversold. The authors marvel at the advances in brain imaging which enable researchers to understand the functions and development of the brain better. However, implications of this research, such as the importance of practice and repetition on the brain, body, and mind, may be more applicable to readers' work. Although the neuroscience clearly supports the authors' concept of higher learning, this chapter may not engage readers unless they have an interest in the science of learning.

Assessment and accountability are addressed in the fifth chapter of the book. Keeling and Hersh believe assessment complements a powerful educational culture and emphasize the importance of assessment in higher education as a way to improve teaching and learning first and as a way to satisfy the calls for accountability second. In particular, formative assessment is touted as a way of providing much needed feedback that improves teaching and learning and allows for remediation and re-teaching so students can attain higher learning.

The authors offer a framework of assessment that is thoughtful and appropriate and links clearly articulated student learning outcomes, pedagogy, appropriate assessment methods, and use of results to improve teaching and learning. This framework should be familiar to anyone engaging in assessment practices, but Keeling and Hersh emphasize that only when the framework is executed as a whole will it truly advance higher learning. It is reassuring that the authors recognize the challenges to developing a strong culture of assessment and they even attempt to neutralize arguments commonly made against assessment efforts.

They begin by acknowledging concerns over the high cost of college and recognize that higher education lacks some efficiency, but indicate clearly that cost is not the real problem – value is. Therefore, the authors focus on increasing value and quality rather than focus on reducing costs.

Keeling and Hersh present their ideas for restoring higher learning to higher education in chapter six. They begin by acknowledging concerns over the high cost of college and recognize that higher education lacks some efficiency, but indicate clearly that cost is not the real problem – value is. Therefore, the authors focus on increasing value and quality rather than focus on reducing costs. This approach is very different from the numerous news stories, speeches, and calls for colleges and universities to adopt a business-like approach to cost-cutting and efficiency. The authors are strongly committed to higher learning and do not believe that incremental changes, mechanistic efficiencies or technological advances are the panacea. For example, they do not approve of replacing full-time, tenure-track faculty with part-time adjunct instructors to reduce salary and benefit costs because of the adverse impact it has on the value of the student experience and ability to achieve higher learning. In no way do they imply that full-time faculty are necessarily better instructors than adjunct faculty, but

rather, full-time faculty have responsibility for mentoring, advising, and engaging students outside of the classroom, all of which contribute to higher learning. Keeling and Hersh do not equate improved technology with higher learning. They stipulate that today's students have greater access to information and computer skills, but point out that many students lack the developmental and higher learning abilities to use, apply, synthesize or critique the information that is available with a keystroke, mouse click or screen tap.

To improve higher learning, colleges and universities need a powerful educational culture or culture of serious teaching and learning where students are viewed as apprentices and can be immersed in rich learning experiences. The authors' description of such a culture is palpable, describing a campus that actually feels different because learning permeates casual conversations, formal lectures, campus activities and cultural events. The authors believe that this culture can and does exist, but acknowledge that the common approach allows for students to sample, not immerse themselves in these activities. The authors' ideal environment is the antithesis of Hollywood's portrayal of college as a place replete with all-night drinking parties, continuous campus festivities, and mischievous coeds.

We're Losing Our Minds is not specific to a residential college environment enrolling mostly traditional-aged students who are full-time students, but it certainly reads that way.

Keeling and Hersh offer a list of ten fundamental principles that when applied intentionally and rigorously will foster a powerful educational culture. These principles include a genuine commitment to student learning, high expectations, clearly articulated standards of excellence, and practices that promote learning and the development of student apprentices from novice to mastery levels. These principles and their recommended applications may be familiar to readers as they are quite similar to the High Impact Practices that foster learning (Kuh, 2008). The authors admit that their suggestions are not new, but most importantly, Keeling and Hersh emphasize the balanced and complete implementation of all of these efforts across campus is required for quality and higher learning. A cherry-picked subset of these practices is insufficient for a transformational impact on students to take place. *We're Losing Our Minds* is not specific to a residential college environment enrolling mostly traditional-aged students who are full-time students, but it certainly reads that way. Professionals at community colleges, urban institutions, or institutions with a large non-traditional aged student body may not connect closely with this book.

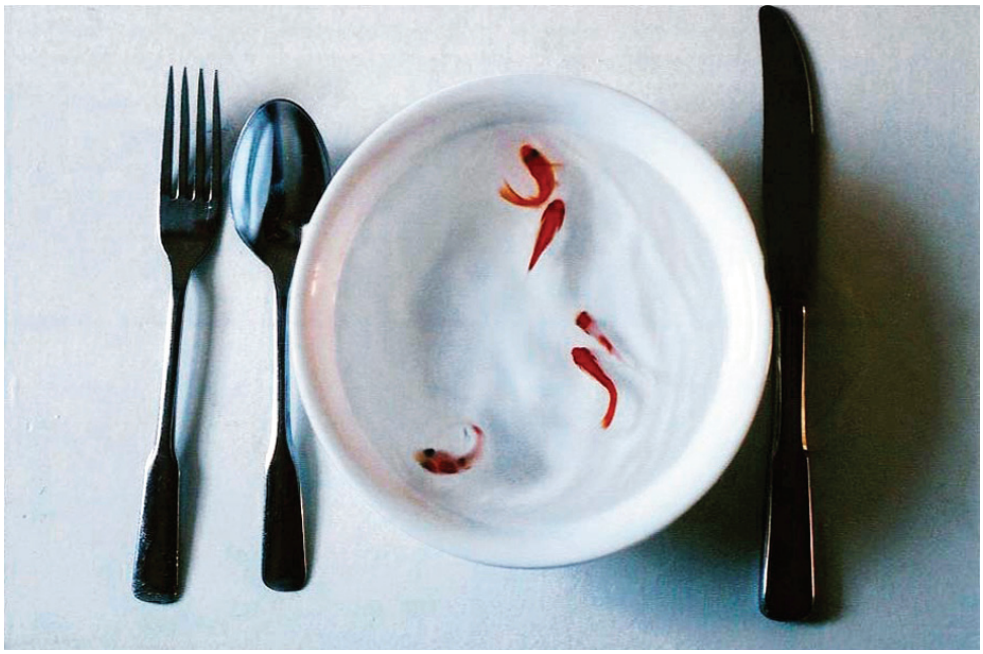
Keeling and Hersh conclude their book in a manner that is seen throughout their work with a summary of the challenges facing higher education and a focus on solutions. The final chapter is a call to action to increase the value of higher education through radical changes and a fully-implemented, holistic approach to student learning and development that immerses a student's body, mind, and spirit in an environment of high standards and expectations focused primarily on higher learning. If Keeling and Hersh are successful in their pursuit to advance higher learning in colleges and universities, stakeholders will likely view college as my father did - the place where one learns how to think.

References

Hebel, S., Stripling, J., & Wilson, R. (2012, June 26). U. of Virginia board votes to reinstate Sullivan. *The Chronicle of Higher Education.* Retrieved from http://chronicle.com/article/U-of-Virginia-Board-Votes-to/132603/

Kuh, G.D. (2008). *High-impact educational practices: What they are, who has access to them, and why they matter.* Washington, DC: Association of American Colleges and Universities.

RUMINATE: INTEGRATING THE ARTS AND ASSESSMENT



NEW COMBINATIONS

Photographer: Huong Fralin
Roanoke, Virginia
www.huongfralin.com

Now, two things are essential for the phenomena incident to the carrying out of such *new combinations*, and for the understanding of the problems involved. In the first place it is not essential to the matter – thought it may happen – that the new combinations should be carried out by the same people who control the productive or commercial process which is to be displaced by the new. On the contrary, new combinations, are as a rule, embodied, as it were, in new firms which generally do not arise out of the old ones but start producing beside them; to keep to the example already chosen, in general it is not the owner of the stage coaches who builds railways.

-*Joseph Schumpeter*

Schumpeter recognized that the knowledge behind the *innovation* need not be new. On the contrary, it may be existing knowledge that has not been utilized before.

-*Michael P. Gallaher, Albert N. Link,* and *Jeffrey E. Petrusa*

Schumpeter's two concepts of combination and resistance, which are independent of one another in his work, can link to each other organically. Under certain conditions, an (existing) combination creates resistance to a new way of doing things. To innovate, in other words, means to break up an existing combination—to break through the resistance it creates, and to replace it with a *new combination*.

-*Richard Swedberg* and *Thorbjørn Knudsen*

GUIDELINES FOR SUBMISSION

All manuscripts submitted to *Research & Practice in Assessment* should be related to various higher education assessment themes, and adopt either an assessment measurement or an assessment policy/foundations framework:

Assessment Measurement:
a) instrument design, b) validity and reliability, c) advanced quantitative design, d) advanced qualitative design

Assessment Policy/Foundations:
a) accreditation, b) best practices, c) social and cultural context, d) historical and philosophical context, e) political and economic context

Article Submissions:

Articles for *Research & Practice in Assessment* should be research-based and include concrete examples of practice and results in higher education assessment. The readers of *Research & Practice in Assessment* are associated with myriad institutional types and have an understanding of basic student learning and assessment practices. Articles for publication will be selected based on their degree of relevance to the journal's mission, compliance with submission criteria, quality of research methods and procedures, and logic of research findings and conclusions. Approximately 40% of submissions are accepted for publication.

Review Submissions:

Reviews (book, media, or software) are significant scholarly contributions to the education literature that evaluate publications in the field. Persons submitting reviews have the responsibility to summarize authors' works in a just and accurate manner. A quality review includes both description and analysis. The description should include a summary of the main argument or purpose and overview of its content, methodology, and theoretical perspective. The analysis of the book should consider how it contrasts to other works in the field and include a discussion of its strengths, weaknesses and implications. Judgments of the work are permitted, but personal attacks or distortions are not acceptable as the purpose of the review is to foster scholarly dialogue amongst members of the assessment community. The *RPA* Editor reserves the right to edit reviews received for publication and to reject or return for revision those that do not adhere to the submission guidelines.

Special Features:

Each issue of *Research & Practice in Assessment* highlights the work of a noted researcher or assessment professional in a manner that aims to extend the scholarly dialogue amongst members of the assessment community. Special Features are invited by the Board of Editors and often address the latest work of the author.

Notes in Brief:

Notes in Brief offer practitioner related content such as commentaries, reports, or professional topics associated with higher education assessment. Submissions should be of interest to the readership of the journal and are permitted to possess an applied focus. The *RPA* Editor reserves the right to edit manuscripts received for publication and to reject or return for revision those that do not adhere to the submission guidelines.

Ruminate:

Ruminate concludes each issue of *Research & Practice in Assessment* and aims to present matters related to educational assessment through artistic medium such as photography, poetry, art, and historiography, among others. Items are encouraged to display interpretive and symbolic properties. Contributions to Ruminate may be submitted electronically as either a Word document or jpg file.

Manuscript format requirements available at: www.RPAjournal.com

RPA BOOKS AVAILABLE

Balzer, W. K. (2010). *Lean higher education: Increasing the value and performance of university processes.* New York, NY: Productivity Press. pp. 312. $51.95 (paper).

Bresciani, M. J., Gardner, M. M., & Hickmott, J. (2010). *Demonstrating student success: A practical guide to outcomes-based assessment of learning and development in student affairs.* Sterling, VA: Stylus Publishing. pp. 224. $27.50 (paper).

Butt, G. (2010). *Making assessment matter.* New York, NY: The Continuum International Publishing Group Ltd. pp. 160. $27.95 (paper).

Cambridge, D. (2010). *Eportfolios for lifelong learning and assessment*. Hoboken, NJ: Wiley, John & Sons, Incorporated. pp. 288. $38.00 (hardcover).

Carey, K., & Schneider, M. (Eds.). (2010). *Accountability in American higher education.* New York, NY: Palgrave Macmillan. pp. 355. $95.00 (hardcover).

Christensen, C., & Eyring, H. (2011). *The innovative university: Changing the DNA of higher education from the inside out.* San Francisco, CA: Jossey Bass. pp. 512. $32.95 (hardcover).

Collins, K.M., & Roberts, D. (2012). *Learning is not a sprint*. Washington, DC: National Association of Student Personnel Administrators. pp. 216. $34.95 (hardcover).

Côté, J. E., & Allahar, A. L. (2011). *Lowering higher education: The rise of corporate universities and the fall of liberal education.* Toronto, ON: University of Toronto Press Publishing. pp. 256. $24.95 (paper).

Dunn, D. S., McCarthy, M. A., Baker, S. C., & Halonen, J. S. (2010). *Using quality benchmarks for assessing and developing undergraduate programs.* San Francisco, CA: Jossey Bass. pp. 384. $45.00 (hardcover).

Flateby, T. L. (Ed.). (2010). *Improving writing and thinking through assessment.* Charlotte, NC: Information Age Publishing. pp. 238. $45.99 (paper).

Hutchings, P., Huber, M. T., & Ciccone, A. (2011). *The scholarship of teaching and learning reconsidered: Institutional integration and impact*. San Francisco, CA: Jossey Bass. pp. 224. $30.00 (paper).

Joughin, G. (Ed.). (2009). *Assessment, learning and judgment in higher education.* New York, NY: Springer Publishing Company. pp. 445. $289.00 (hardcover).

Makela, J.P., & Rooney, G.S. (2012). *Learning outcomes assessment step-by-step: Enhancing evidence-based practice in career services*. Broken Arrow, OK: National Career Development Association. $35.00 (paper).

Maki, P. L. (2010). *Assessing for learning: Building a sustainable commitment across the institution*. Sterling, VA: Stylus Publishing. pp. 356. $32.50 (paper).

Martin, R. (2011). *Under new management: Universities, administrative labor, and the professional turn.* Philadelphia, PA: Temple University Press. pp. 253. $69.50 (hardcover).

Noyce, P. E., & Hickey, D. T. (Eds.). (2011). *New frontiers in formative assessment*. Cambridge, MA: Harvard Education Press. pp. 260. $29.95 (paper).

Vande Berg, M., Paige, R. M., & Lou, K. H. (2012). *Student learning abroad: What our students are learning, what they're not, and what we can do about it.* Sterling, VA: Stylus Publishing. pp. 470. $39.95 (paper).