Ubiquity Symposium

MOOCs and Technology to Advance Learning and Learning Research

Assessment in Digital At-Scale Learning Environments

by Piotr Mitros, Anant Agarwal, and Vik Paruchuri

Editor's Introduction

Assessment in traditional courses has been limited to either instructor grading, or problems that lend themselves well to relatively simple automation, such as multiple-choice bubble exams. Progress in educational technology, combined with economies of scale, allows us to radically increase both the depth and the accuracy of our measurements of what students learn. Increasingly, we can give rapid, individualized feedback for a wide range of problems, including engineering design problems and free-form text answers, as well as provide rich analytics that can be used to improve both teaching and learning. Data science and integration of data from disparate sources allows for increasingly inexpensive and accurate micro-assessments, such as those of open-ended textual responses, as well as estimation of higher-level skills that lead to long-term student success.

Ubiquity Symposium

MOOCs and Technology to Advance Learning and Learning Research

Assessment in Digital At-Scale Learning Environments

by Piotr Mitros, Anant Agarwal, and Vik Paruchuri

Assessments in courses serve several purposes. Throughout the course, assessments provide a way of monitoring what students know, which allows instructors and students to tailor teaching and learning to problematic areas (initial and formative assessment) [1]. Assessments act as one of the principal means of student learning in the role of providing a means both to practice skills and receive feedback [2], as well as deriving or constructing knowledge [3]. Following the course, assessments provide a measure of how well an individual student performed, as well as the effectiveness of the course (summative assessment). They are also one of the key components of grading, which is typically structured with a combination of goals, including certifying student accomplishment and providing motivation.

Students and instructors are incentivized to optimize teaching and learning to testable skills, sometimes at the expense of more difficult-to-test skills. As a result, limited or inaccurate assessments can actually harm teaching and learning. Assessment in traditional education is tremendously resource-intensive, which severely restricts what can be assessed. This problem is particularly true in high-stakes tests, such as SATs, where exams are typically three to four hours long, and must be graded for as millions of students in bulk. In most cases, high-stakes exams measure simple skills and use those as proxies for more complex skills—such as mathematical maturity, critical thinking, and complex problem solving—but completely fail to capture interpersonal skills such as team work. While time constraints in traditional classroom settings are somewhat more relaxed than in high-stakes exams, instructors still often rely on proxies for complex skills. For example, when measuring communications, the most common proxy is an essay—a medium relatively rare in the workplace. Instructors cannot effectively

critique longer formats of communications, such as email threads, meetings, and similar without extreme student to faculty ratios.

Digital at-scale learning environments—defined as learning environments that accommodate thousands of students, either in the format of MOOCS or as common course shared across many classrooms—employ several tools to help address these issues. In particular, centralization allows more resources to be invested in the design of assessments, allowing for improved technologically and pedagogically clever assessments whose design is driven by what we know about learning. Combining multiple data sources including human insight, simple machine heuristics, and artificial intelligence allows us to assess more complex skills.

Machine-assisted Tools for Open-Ended Responses and Feedback

Digital assessments have long been effective as a technique to liberate instructor time, which in turn puts more focus on complex skills, particularly in blended learning settings, and provide immediate formative feedback allowing students and instructors to focus on problem areas [4]. Building on this work, we are increasingly seeing approaches where humans and machines work in concert to more quickly and accurately assess and provide feedback to student problems.

Systems such as Microsoft PowerGrading [5] and MIT Caesar [6] provide tools by which instructors can provide feedback to common problems, and apply that feedback to a large number of students experiencing the same issue. Discussion forums such as Piazza, Askbot [7], and edX serve two roles. The more obvious role is to give students a means to ask questions and receive responses. Although such student interaction leads to substantial gains in learning, in courses we have examined, only a minority of students actively ask and answer questions. Likely, the more important role is once a question is answered that feedback is collected, and using relatively simple search tools other students with the same problem can find and reuse that feedback. In addition, the system can track which feedback was useful to which students, providing instructors with knowledge about common misconceptions. Since in all three of these cases, creation of feedback for common problems is repeated less often. This class of machine-assisted techniques allows students to receive feedback from assessment at much higher levels of personalization, quality, and quantity for the same level of human effort.

The open-source edX Open-ended Response Assessment (ORA) system builds on this work by blending four different sources of assessment and feedback:

- **Self-assessment** has students rate their own answers on a rubric. Given a clear rubric and a way to disincentive students from cheating, self-assessment can provide very accurate and immediate feedback.
- **Peer assessment** has students provide grading and feedback for assignments submitted by other students. Given a clear rubric and training in how to grade, calibrated peer assessment can also provide results of accuracy comparable to expert instructors [8]. In the edX ORA system, students practice grading sample problems with known grades prior to grading other students. It is still an open question as to how to best design peer assessment for assessments without a clear rubric (such as general grading of writing quality), although a number of approaches have been tried with varying levels of success. In environments with many students simultaneously online, peer assessment can provide rapid (although not immediate) feedback.
- **Instructor assessment** is considered the gold standard for quality of grading and feedback. It is very expensive with large numbers of students, and provides relatively slow feedback.
- **AI assessment** has a computer grade essays by attempting to apply criteria learned from a set of human-graded answers. AI assessment can provide immediate feedback, and is very accurate for short, factual answers, as well on some (but not all) aspects of more complex assignments. If the problem rubric includes common student errors, AI assessment can provide qualitative feedback for those errors. Depending on how it is used, AI assessment may be gameable by students [9].

Each source assesses on a rubric that consists of a set of axes on which the student is evaluated. These may be well-defined criteria (E.g. “Did the student mention the concept of conservation of energy?”) or qualitative (e.g. “On a scale of 1-5, rate the quality of the logic of the argument.”). In addition, the systems may return qualitative text feedback, as well as a certainty estimate.

ORA integrates these four approaches in order to maximize the speed, quality, and accuracy of assessment and feedback for a given level of human effort. In the theoretical formulation [10], each of the four grading systems contributes a different type and amount of information. The system routes problems to the most appropriate set of grading techniques. An algorithm combines responses from graders to individual rubric items into feedback and a final score. The

current implementation uses predefined static flows, where the course creator defines how a problem is graded. For example, an instructor may ask students to do self- and peer-review prior to giving instructor feedback, or may combine peer-review with AI assessment to leverage the relative strengths of both. As we will explore in the next section, self-assessment alone can be an especially powerful technique.

Active and Mastery Learning

The edX platform is based on mastery learning [11], active learning, and tools with which students actively monitor their level of expertise. Mastery learning is a technique where students do not move on from a concept until after they have mastered it to a sufficient level to tackle future material. In a broad range of studies, mastery learning has been shown to give very substantial gains in learning [12] Metacognition—in this context, the ability to monitor one's knowledge and learning processes—has also been shown to give substantial gains in learning in a broad range of experiments, although in contexts very different from edX [13]. Whether we see similar gains is still an open question.

The first edX course, Circuits and Electronics (MIT 6.002x), implemented mastery learning through the use of entirely machine-gradable open-ended responses, such as circuit schematics (verified by simulation), equations (verified by sampling), and numbers. Since the answer to these types of questions cannot be guessed, students could attempt to submit an answer as many times as necessary in order to understand and solve a problem correctly [7]. In courses that do not lend themselves to this type of open-ended assessment, mastery learning typically uses a large problem bank from which students must complete some number of questions correctly in succession before moving on.

During knowledge delivery, 6.002x allowed students to self-monitor their level of mastery by interleaving assessments with text and videos (such assessments were not counted as part of the student's grade). If a student failed to master a portion of a video, they could re-watch it and retry the assessment. With ungraded problems, students have no incentive to game the system, so both self-assessment and AI grading can be very helpful even in isolation. However, since self-assessment gives away the answer, giving students multiple tries would require a problem bank.

Open Learning Analytics and Complex Skills

Aside from assessing individual problems, digital at-scale learning environments can enable more accurate evaluations of a student's ability to master complex skills by integrating data from multiple sources. Common systems that do this, such as Purdue Course Signals [14], Marist Open Academic Analytics Initiative [15], and Desire2Learn Student Success System [16] attempt to assess a fairly high-level but amorphous measure of skill—an estimate of how likely a student is to succeed in a given course. As this technology improves, this is likely to move into assessment of more nuanced skills, such as teamwork, organizational skills, and time management.

Open analytics architectures [17], such as edX Insights and Tin Can, provide frameworks for this type of integration by providing a common data repository and a set of APIs to access that data. Fully integrated into a learning environment, such frameworks can virtually monitor all learning interactions that a student might engage in digitally. Once these frameworks are deployed, they will allow rapid prototyping of assessments of more complex competencies, as well as more accurate assessments of simple competencies.

Natural language processing frameworks, such as the open-source edX EASE and Discern, can potentially monitor student interactions over chat, forums, or emails, and begin to give insights into students' soft skills, writing process [18], communications styles, and group dynamics. In addition, integrated metrics such as project-based assignments across multiple courses, comparing group performance to individual performance, have the potential to give new measures of complex, previously difficult-to-measure competencies.

For simple competencies, techniques such as item response theory [19] can help in the calibration of difficulty of problems. Once calibrated, they allow integration of a student's performance throughout a course with that student's performance on a high-stakes assessment to give a more accurate estimate of student ability.

Conclusion

While many of the goals of an educational experience cannot be easily measured, it is much easier to improve, control, and understand those that can. Tools for analytics and assessment, such as edX ORA, Insights, EASE, and Discern, provide several new means to assess students in

ways which can improve the depth, frequency, and response time. Potentially expanding the scope with which students and instructors can monitor learning, including assessment of higher-level skills, and proving personalized feedback based on those assessments.

References

[1] Sadler, D. R. Formative assessment and the design of instructional systems. *Instructional science* 18, 2 (1989), 119-144.

[2] Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. The role of deliberate practice in the acquisition of expert performance. *Psychological Review* 100, 3 (1993), 363.

[3] Chi, M. T. H. The ICAP Module: Guidelines for teachers to increase students' engagement with learning. Unpublished funded proposal to IES. Arizona State University, Tempe. 2011.
[4] Novak, G., Patterson, E., Garvin, A., Christian, W. *Just-in-Time Teaching: Blending Active Learning with Web Technology.* Prentice Hill, Upper Saddle River, NJ, 1999.

[5] Basu, S., Jacobs, C., and Vanderwende, L. Powergrading: A clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics (TACL)* 1 (2013), 391−402.

[6] Tang, M. Caesar: A Social Code Review Tool for Programming Education. M.Eng. Thesis. Massachusetts Institute of Technology, 2011.

[7] Mitros, P. F., et al. Teaching Electronic Circuits Online: Lessons from MITx's 6.002 x on edX. *2013 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, New York, 2013.

[8] Walvoord, M E., et al. An Analysis of Calibrated Peer Review (CPR) in a Science Lecture Classroom. *Journal of College Science Teaching* 37,4 (2008), 66.

[9] Winerip, M. Facing a Robo-Grader? Just Keep Obfuscating Mellifluously. *New York Times*. April 22, 2012. Accessed April 24, 2014.

[10] Mitros, P., Paruchuri, V., Rogosic, J., and Huang, D. An Integrated Framework for the Grading of Freeform Responses. *MIT Learning International Networks Consortium.* June 17, 2013.

[11] Bloom, B. S. The 2 Sigma Problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher* 13, 6 (1984), 4-16.

[12] VanLehn, K. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist* 46, 4 (2011), 197-221.

[13] National Research Council. *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. The National Academies Press, Washington, DC, 2000, 67-68, 97-98

[14] Arnold, K. E., and Pistilli, M. D. Course Signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the Second International Conference on Learning Analytics and Knowledge*. ACM, New York, 2012.

[15] Lauría, E. J. M., Moody, E. W., Jayaprakash, S. M., Jonnalagadda, N., and Baron, J. D. (2013). Open academic analytics initiative. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK '13* (p. 150). ACM, New York, 2013. doi:10.1145/2460296.2460325.

[16] Essa, A., and Ayad, H. Student Success System: Risk analytics and data visualization using ensembles of predictive models. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. ACM, New York, 2012.

[17] Siemens, George, et al. Open Learning Analytics: An integrated and modularized platform. Society for Learning Analytics Research (SoLAR). 2011.

[18] Southavilay, V., et al. Analysis of Collaborative Writing Processes Using Revision Maps and Probabilistic Topic Models. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. ACM, New York, 2013.

[19] Embretson, S. E., and Reise, S. P. *Item Response Theory*. Psychology Press, 2000.

About the Authors

Anant Agarwal is the president of edX, the online learning destination founded by Harvard and MIT. He taught the first edX course on circuits and electronics. He has served as the director of MIT's Computer Science and Artificial Intelligence Laboratory, and is a professor of electrical engineering and computer science at MIT.

Piotr Mitros is the Chief Scientist of edX. As a Research Scientist at MIT, Mitros lead the technical and pedagogical development of the MITx platform, which later evolved into the edX platform. Mitros holds Ph.D. (2007), M.Eng. (2004), and B.S. (2004) degrees, all from MIT.

Vik Paruchuri is a software developer living in Cambridge, MA. He received his B.A. in American History from the University of Maryland. He has worked on automated essay scoring, machine learning, web projects, and mobile development. He placed first in the Hewlett Foundation Short Answer Scoring Competition in September 2012.

DOI: 10.1145/2591795

