CHAPTER 8 – Assessing Individual Learner Performance in MOOCs

Ryan S. Baker[1], Piotr Mitros[2], Benjamin Goldberg[3], and Robert A. Sottilare[3]
University of Pennsylvania[1], edX[2], US Army Research Laboratory[3]

Introduction

Massive open online courses (MOOCs) have emerged as a prominent mode of online education, but the quality of assessments in MOOCs remains inconsistent. There has been a consistent gap between the state of the art in assessment and the state of the practice in both MOOCs and other forms of educational technology. Improving the quality of assessment has the potential to improve their usefulness for certification, formative feedback, and learning. In this chapter, we discuss this gap and efforts to improve assessment in MOOCs. There are several potential directions for improving assessments in MOOCs, including improving the psychometric properties of simple assessment types, such as multiple-choice and fill-in-the-blank questions; creating richer assessment experience through more student interaction and more engaging experiences (e.g., games, and simulations), through leveraging help resources to see how students perform with some degree of support; and using more powerful technology such as automated essay scoring to better assess students. We discuss both existing efforts and ways that research in other communities can be incorporated into MOOC platforms.

In fall 2011, a small group of computer scientists from Stanford University launched the first xMOOC (an experimental MOOC based on a traditional university course), the Stanford AI Course. This early MOOC platform only supported three assessment types – multiple choice questions, select-many multiple choice questions, and numeric answers. However, the course used these types of assessment in a relatively sophisticated way. Some assessments were embedded in the video and tightly integrated with content presentation, giving continuous formative assessment throughout the course. As instructors walked through mathematical derivations, students would do portions of the work (this practice is sometimes referred to as a partially faded worked example – e.g., Salden et al., 2010). In other assessments, students were given a problem, asked to develop an algorithm to solve that problem, and asked to enter the output of their program implementing that algorithm into a numeric response question. This resulted in a relatively rich student experience.

Shortly after the Stanford course, Coursera introduced two courses, one in machine learning and the second in databases. Coursera's initial courses relied more on simple multiple choice questions, although they introduced problem banks to create the possibility of mastery learning through selecting items from the problem banks and continuing to offer items until the student demonstrated mastery. In addition, Coursera introduced assignments that could be graded via external software, a technology mostly used for programming assignments.

The MITx platform, now Open edX, was introduced approximately 6 months later. In the first MITx course, students worked through design and analysis problems which they could answer with numbers, equations, and circuit schematics. Their answers were verified with circuit simulations (in a JavaScript Spice clone), Python programs, plugging numbers into equations, or comparison of numbers with tolerance. Even with numeric answers, there were often multiple correct answers. For example, a design problem that required students to pick component values for a resistive divider of a given attenuation would need to verify 1) ratios of those components, 2) whether they were valid values, and 3) whether they were realistic values (Mitros et al., 2013). This type of nuanced complexity to student solutions is not unique to this course, but is found in a range of higher education course assessments, particularly in domains such as engineering.

The first dozen edX courses launched with several assessments that were discipline-specific, such as tools for evaluating computer code, chemical equations, and visualizing crystallography planes.

Around the same time, Coursera introduced calibrated peer assessments where an essay or other artifact is graded by a student's peers, allowing for human assessment at a scale that would not be feasible for an instructor or teaching assistant in a course with tens of thousands of students.

However, despite the availability of support for programmatic auto-grading, calibrated peer assessment, and other forms of rich assessment embedded into teaching activities, most MOOCs rely upon relatively simple forms of assessment. The majority of MOOCs rely upon weekly quizzes or assignments where the student answers a fixed set of multiple-choice or fill-in-the-blank questions. While edX has support for over 50 different types of activities available and ready to use as options for course authors, around 80% of assessments are either multiple choice questions or similar simple assessments. Most competing platforms rely even more heavily on multiple-choice questions.

When compared to the rich and broad range of assessment seen within computer-based learning environments, for example, especially within the research community, the limitations of current practice in MOOCs and many other digital learning technologies, become especially notable. While learners using computer-based learning environments such as Cognitive Tutor are assessed at the same time they are learning (Corbett & Anderson, 1995), assessment in MOOCs is often separate from learning. For example, although edX is designed around interleaved learning sequences with integrated formative assessment, only half of the learning sequences with videos in courses deployed on edX.org include integrated assessments. While adaptive assessments in systems like the Assessment and Learning in Knowledge Spaces (ALEKS) determine where a student is in the progression from learning prerequisites to learning the skills that build on those prerequisites (doignon & Falmagne, 2012), MOOC developers rarely have the resources to create sufficient numbers of assessments to enable adaptive pathways, especially given the thousands of courses used at a university level. Consequently, adaptivity is limited to feedback and hinting. While learners using simulations take part in rich performance assessments (Clarke-Midura et al., 2012), MOOC learners often are assessed solely on multiple choice and fill-in-the-blank items, since such assessments have not yet been developed for the diversity of topics covered in MOOCs. Many examples of richer assessment can be found. Applying these approaches to the scale and diversity of courses offered as MOOCs remains an open challenge. The problem is complicated by the types of instructors teaching MOOCs – in contrast to trained educators at a K–12 level, university courses (and the MOOCs based on them) are generally taught by subject-matter experts, with little background in teaching and learning or educational technology.

In this chapter, we discuss the efforts that have been made to enhance assessment within MOOCs and potentials for enhancing assessment in MOOCs further. As MOOCs continue to serve tens of millions of students with thousands of courses from hundreds of universities, enhancing assessment in this context has the potential to improve learning experiences for many people in many contexts.

Improving the Quality of Current Common Types of Assessments

As mentioned previously, many MOOCs assess learners with a fixed set of multiple-choice and fill-in-the-blank items. Even within the paradigm of this type of item, there are several limitations to current practice in this regard. One limitation is that items are typically not validated. Item validity in assessment construction is a well-known area, incorporating both the psychometric properties of items (Dudycha & Carpenter, 1973) and their mapping to the content goals for the assessment (Mislevy & Riconscente, 2006). However, applying such algorithms in MOOC settings is still an area of open research (Champaign et al., 2014). There are several challenges to integrating such algorithms into MOOCs:

- When courses rely on mastery learning, students have much more incentive to guess and explore. Incorrect answers are often a result of such exploration, rather than item difficulty, as assumed in models such as the Rasch model or other Item Response Theory (IRT) models.

- MOOCs are open, and many learners have no intention of completion. Many students only target sections of courses that are relevant to them. More advanced learners are likely to skip over or rush through easy items, while novices might skip difficult problems, which can reduce the accuracy of knowledge estimation.

- In contrast to an exam where knowledge stays constant during the course of the exam, or adaptive systems that alternate between learning and assessment (such as ALEKS), students in MOOCs learn over time, and items often have intellectual cohesion and sequential narrative. Models which presume learner knowledge is fixed, such as IRT, fail to work in such cases. Another popular knowledge modeling algorithm, Bayesian knowledge tracing (BKT), assumes that learners learn, but assumes a certain level of independence in the order of items. Learners self-regulate when working through MOOCs, invalidating such assumptions.

- MOOC assessments rarely have a 1-1 mapping to learning components, unlike psychometrically designed tests. Analysis and design courses will have complex, multi-concept questions, where a single problem might take several hours to complete (Seaton et al., 2014), while history courses might spot-check facts, such as asking about a key date. Many psychometric models are underconstrained in such cases.

- MOOC offerings typically decline in size with each run, where the first run is substantially larger than successive ones. Applying data from the first run to improve future runs would not impact the large number of students in the first run, so improvements would either need to be developed based on a prototype cohort[1], or run in real time by analyzing and attempting to enhance problems after a minority of students attempted them (as found in systems such as the Learning Online Network with Computer-Assisted Personalized Approach [LON-CAPA]).

Another key difference is that MOOC assessments typically grade students with regard to percentage correct. Psychometricians have been aware for decades that items have different properties – even in the same topic area, items may have different difficulties and different degrees of discrimination (how effectively they distinguish skillful students from non-skillful students), and as such, grading according to percentage correct is far from optimal for measurement of student skill. However, psychometric considerations must be balanced with other goals of exercises in courses:

- Exercises serve as a primary means of learning. Mastery learning is a key technique for effective learning, repeatedly shown to lead to better learning outcomes (Bloom, 1987). In mastery learning settings, students are asked to continue to attempt problems until they demonstrate mastery, for example by getting a sufficient number of problems correct or according to the assessments of knowledge modeling algorithms (e.g., Anderson et al., 1995). This is, in many ways, opposite to the goals of measurement, where items students answer correctly 50% of the time contribute maximum information. A question that all students answer correctly is often an excellent tool for learning, but contributes little psychometric information.

[1] Many MOOCs first run in a residential format, and some MOOC providers have experimented with a prototype cohort with fewer students for validation of courses.

- Grading is used to motivate students. For those purposes, simplicity is paramount; students ought to understand grading schemes. Algorithms such as multivariate IRT are opaque to students.

- Grading is used for certification. In this context, it is important for users of such accreditation to understand what the accreditation means.

Nonetheless, MOOCs give a new opportunity for grading schemes which integrate psychometric considerations into learning engineering and learning design. Some pilot work has attempted to take these several issues into consideration within the context of MOOCs – for example, Colvin et al. (2014) uses a form of IRT to better assess students in a physics MOOC, taking item difficulty and discrimination into account. Pardos and colleagues (2013) similarly extend BKT to assess the degree of student learning over time in a MOOC on circuit design, taking item difficulty and multiple student attempts into account. However, these innovations have not scaled to the broader range of MOOCs or led to better achievement of the non-psychometric goals of exercises and assessments in MOOCs. How to do so effectively, what issues will come up with the diverse range of courses and pedagogies in MOOCs, and how to balance the competing design goals of assessments all remain open problems.

These problems are unlikely to be solved until tools that ingest MOOC data are standardized and widely deployed. Instructors and course development teams generally do not have the capacity to do this type of research and enhancement on their own. There are a number of initiatives to create such tools (Cobos et al., 2016; Dernoncourt, 2013; Fredericks, 2016), but none are standard or complete. A widely useful platform for open, integrated learning analytics (Siemens et al., 2011) remains a dream, for both technical and non-technical reasons. Scaling existing learning algorithms to perform the complex calculations needed for classical psychometric modeling on the terabytes of data in MOOCs in real time remains a technical challenge, while sharing and integration of proprietary learning data are open legal and policy questions.

Broadening Assessment in MOOCs

Another way that MOOC assessment could be enhanced is through broadening formats for assessment. As noted previously, the first xMOOC alternated between providing conceptual content in the form of video lectures, having students manipulate information, and assessing student understanding. Many xMOOCs have included rich analysis-and-design problems. While the functionality to do this remains present in the primary xMOOC platforms, it is underutilized; at least half of MOOCs do not make adequate use of such functionality[2].

A common strategy in K–12 intelligent tutoring systems (ITSs) is to build problem sets where a larger problem is broken down into different steps, and the student receives feedback at each step. This is broadly the strategy used in the highly successful mathematics ITS, Cognitive Tutor (Anderson et al., 1995). An example of these step-based intelligent tutoring problems are seen in Aleven et al. (2015), where a data science MOOC's assessments were converted into step-based problems. This allows more frequent feedback to students, and better tools for understanding student learning and knowledge gaps. However, it does not easily apply to all problems that might be seen in MOOCs. Many university courses strive for more complex, multi-concept authentic assessments, where students must not only work through calculations, but come up with a high-level problem-solving strategy in an open-ended setting. In an intermediate approach, students work in the platform in open-ended tools such as word processors, circuit schematic entry tools, or code editors, and the platform monitors student work. Traces of such activities mined for data.

[2] Approximately half of learning sequences on edX.org with videos do not include assessments. This number may be even lower on other MOOC platforms.

However, such systems are extremely expensive to build since such analyses tend to be very domain-specific. This engineering cost would be prohibitively expensive for the thousands of courses offered at a tertiary level.

Additional types of problems may afford richer experiences, both in terms of learning and assessment. For example, in the ITS Betty's Brain, students create concept maps to explain the interrelationships in a domain, and then the concept maps are evaluated in terms of their match to an expert-generated concept map (Leelawong & Biswas, 2008). Bringing such functionality into MOOCs is complex. Several MOOCs have encouraged learners to generate concept maps (e.g., Viswanathan, 2012; Bachelet et al., 2015), but MOOC developers have not yet leveraged the opportunity to automatically assess and provide immediate feedback on the resultant concept maps. There are several issues with bringing such tools to scale:

- Such tools tend to be complex to develop, both technically and pedagogically. University courses are taught by subject matter experts, often with little background in teaching-and-learning, educational technology, or computer programming. Although edX.org has over 50 activity types available to course authors, the Open edX ecosystem has at least twice that number, and many more are integrated through Learning Tools Interoperability (LTI), iframes, or JavaScript, only a minority of course teams are able to make use of such functionality.

- Developing such tools is expensive. In the tertiary space, there are thousands of courses[3], each taken by thousands or tens of thousands of students. Developing custom technology for each of these would cost hundreds of millions of dollars.

- Even seemingly broadly applicable tools, such as the concept maps mentioned, only apply to a minority of courses. Many university courses cover areas of active research, and as such have poorly-defined concepts and learning objectives. In these courses, the curriculum and objectives are still being defined.

A promising area of research is finding and encouraging the use of simple tools – both from a developer and course author standpoint – to enable cognitively complex tasks such as the concept maps seen in Betty's Brain. For example, in the context of learning-by-teaching, MOOCs have used Q&A forums, peer feedback tools, and community TAs. With thousands of mature students in a course, it is often possible to come up with relatively simple techniques which mirror the cognitive processes of students in ITSs, but do so by relying on either the intelligence of crowds or the intelligence of individual students.

Perhaps even richer interactivity and assessment can be found in systems that allow students to enter answers to conceptual questions in natural language, such as AutoTutor (Graesser et al., 2005). AutoTutor uses natural language processing (NLP) both to evaluate student responses and to ask probing questions that help to explore how much students understand. This type of assessment, while expensive to create, can help to richly explore student understanding, toward offering more sensitive responses and support to students. However, it is relevant to ask how to build out this type of learning and assessment activity in an

[3] According the National Center for Educational Statistics, even large majors such as computer science only have on the order of 40,000 graduates every year. Smaller majors might have single-digit thousands. A high-caliber but narrow school such as the Massachusetts's Institute of Technology (MIT) offers roughly 2,000 courses to cover primarily science and engineering education. If we assume similar numbers of courses across disciplines MIT does not offer, such as agriculture, education, medicine, or law, a complete set of courses to cover a broad university education would require about 10,000 courses. As of 2017, there are thousands of MOOCs.

economical fashion for the number of courses in MOOCs. It is also uncertain how systems like AutoTutor will integrate into the design of current MOOCs.

The written word can also be the focus of MOOC assessment through the grading of essays and other extended written work. Several MOOCs have used peer review to grade essays, where (as mentioned previously) students grade each others' work and give feedback (Balfour, 2013). However, peer review may be less useful for problems with a large expert-novice gap, where a substantial portion of the goals may be, for instance, to give feedback on how well students conform to good design practice or other professional conventions. In addition, many students prefer to have their assignments graded by an expert rather than a peer (who may even be less knowledgeable than they themselves are) (Luo et al., 2014). However, in the majority of cases, with a clear, well-designed, and relatively closed-ended rubric, and proper calibration, peer reviews can be more reliable than a single expert review for suitable assessments – and are definitely more scalable.

Another potential approach to grading student essays is to use auto-grading, based on NLP. NLP-based auto-grading has been used at scale in other domains, perhaps most notably in standardized examinations. A thorough review of key systems for automated essay scoring can be found in (Dikli, 2006). Reilly and colleagues (2014) report on the use of automated essay scoring in a pharmacy MOOC, finding good agreement between automated scoring and instructor scores. However, automated essay scoring remains controversial in the context of MOOCs. edX piloted this capacity and found that when appropriately used, it had very high quality results (Mitros et al., 2013). However, practical engineering constraints prevented this approach from out to large numbers of courses. In particular, for such algorithms to work, the instructor must first hand-grade around 100 submissions. These free-form text submissions are typically unavailable the first time a MOOC is run, creating a chicken-and-egg problem. In addition, while the system worked very well in the courses in which it was piloted it is difficult to predict whether it would work in all such contexts. If an auto-grading system failed in a MOOC, leaving thousands of essays to be hand-graded, such a failure could either be exceptionally expensive or highly problematic. Such problems are solvable, whether by integrating with peer grading, use in prototype courses to obtain initial models, using teaching assistants in the developing world, or other solutions, but no solutions has currently been developed to the level of being production-ready for thousands of courses.

Across approaches to assessment and instruction, one key lesson learned is that in MOOCs, usage of a feature is strongly tied to how easy it is for course authors to discover and use that feature. Course authors are willing to invest significant effort into making high-quality courses if they can figure out how to do so. For example, ITSs have a range of hinting functionality, including on-demand hints and so-called "bug messages" for incorrect answers. Such hints both enhance learning and have been shown to be a valuable component of more precise assessment, providing data beyond just student correctness that is predictive of long-term outcomes (Feng, Heffernan & Koedinger, 2009). The edX platform added simple authoring functionality for hinting in 2015, including hints in default template problems. As a result, as of this writing, around ⅔ of edX assessments have hint functionality, either as on-demand hints or "bug messages".

Assessment: Beyond Knowledge

Thus far in this chapter, we have discussed the assessment of students in MOOCs as if the only thing worth assessing is students' knowledge and skills. It is true that knowledge and skill have historically been the primary focus of assessment work, across contexts and domains, but they are hardly the only constructs that can be assessed, or the only constructs that should be assessed. MOOCs are well suited to assessing more complex skills, such as group work, creative problem solving, and leadership. MOOCs capture minute click-by-click student interactions, across a diverse range of subjects, with data for some learners across up to five years. These data have the potential to help us study and assess student progress across a sequence of several complex and group projects, providing insights into the details of the social interactions and problem solving within those projects (Mitros et al., 2014). However, MOOCs have not yet reached their full potential in this area.

For instance, one of the key areas of research in the ITS community over the last decade has been the assessment of metacognition and self-regulated learning (SRL) skill. MOOC students are generally gifted high school students, college students, and adult learners, and even within that set, are disproportionately autodidacts. They can be hypothesized to have higher levels of SRL skills and metacognition than the more general population of learners whom ITSs traditionally target. The edX platform, while offering a linear default path through content, was designed to support SRL strategies by providing supplementary resources that students can choose to access, as well as multiple navigational elements for students to be able to monitor their learning, skip over material they know, or navigate back to material they did not adequately master[4]. However, evaluation of how well these design elements work is limited, and it is not yet known to what degree the edX design works as intended. In terms of assessing metacognition and SRL within MOOCs, some work has focused on the use of out-of-context questionnaires rather than recognizing SRL from behavior (Hood et al., 2015; Onah & Sinclair, 2016). However, it has been argued that this type of questionnaire does not capture key aspects of SRL (Winne & Baker, 2013). Other work has looked at whether students' navigation patterns in MOOCs follow the default linear path, but has not fully closed the loop from quantitative description to qualitative understanding (Guo et al., 2014). We do see that a significant number of students completing MOOCs skip over significant numbers of videos (Seaton et al., 2014), which suggests some different learning strategies are being applied.

In the context of ITSs, models have been developed that can recognize a range of SRL strategies, from unscaffolded self-explanation (Shih, Koedinger & Scheines, 2011) to help-seeking (Aleven et al., 2006). Models that can assess help-seeking skills have been used as the basis of automated support for SRL, leading to systems that produce enduring positive changes in students' help-seeking strategies (Roll et al., 2011). Even simple training in strategies for planning, monitoring, and knowledge elaboration can lead to better learning outcomes in laboratory studies (Azevedo & Cromley, 2004). However, simply recommending SRL strategies to MOOC learners does not appear to lead to benefits (Kizilcec et al., 2016). Overall, the best way of encouraging effective SRL strategies in MOOCs is an open question.

MOOCs offer several opportunities for measuring SRL at a behavioral level, including student use of discussion forums to use questions, and student activity in the face of incorrect answers within a knowledge assessment – after making a mistake, does the student give up? Try again immediately? Ask for help (perhaps on the discussion forum)? Re-watch the video? By modeling these behaviors, we may be able to assess SRL in MOOCs in the same rich fashion as has been achieved for ITSs.

Another area of assessment in MOOCs that goes beyond knowledge and skills, and which has received relatively more attention, is the assessment of student engagement. Inspired by student success systems in for-credit programs (Arnold & Pistilli, 2012), many researchers have attempted to identify the factors associated with a student not completing a MOOC and predict in advance whether or not (and when) a student will stop participating in a MOOC (Jiang et al., 2014; Kloft et al., 2014; Yang et al., 2013; Sharkey et al.,

[4] 6.002x, the original edX course, was structured as learning sequences, composed of roughly a dozen elements each, replacing what would be lectures in traditional classrooms. These was linear navigation with back/forward buttons, but also a set of icons, one for each element of the sequence, with the icon indicating the type of element (e.g., problem vs. video), and tooltips describing what each element is about. Students, anecdotally, took multiple strategies. For example, some students would navigate to assessments and only watch videos if they had problems with those assessments. Videos had links to multiple additional means of presentation, such as textbook pages. Within the video, multiple speeds were available, for moving through the video more quickly or more slowly. In addition, a scrolling transcript allows students to read ahead and navigate to precise points in a video.

2014). Across studies, it appears that several forms of participation are associated with MOOC completion, including posting to discussion forums, reading discussion forums, completing assignments (somewhat tautologically, since completion of a MOOC is typically based on completing assignments with a sufficiently high grade), and watching videos (see review in Andres et al., in press).

A related area of research is in attempting to infer MOOC learners' emotions or sentiment. This is a well-established area in other types of online learning system (see the review in Baker & Ocumpaugh, 2014), with researchers developing, validating, and using models of affective states such as boredom, frustration, and engaged concentration/flow. Researchers have begun studying assessment of sentiment in MOOCs as well. For instance, Wen and colleagues (2014) use discussion forums data on Coursera to determine if students have positive or negative attitudes to a course's lectures, assignments, and peer assessments. Chaplot and colleagues (2015) show that student sentiment from discussion forums can be incorporated into models that infer whether a student will drop out of a course, leading to more accurate prediction of retention. However, there is not yet work to detect emotions in MOOCs beyond simply positive and negative sentiment; research into more complex emotion in MOOCs has thus far depended on self-report instruments (Dillon et al., 2016) rather than the automated detectors used with ITSs and other types of artificially intelligent learning software.

Finally, some researchers have begun to study how behavior within MOOCs can be predictive not just of completing the MOOC, but of students' career trajectories after the MOOC. Chen and colleagues (2016) find that many students learning in programming MOOCs take their knowledge beyond the MOOC, incorporating new programming knowledge into their publicly released software on gitHub. Wang and colleagues (2017) have determined that reading discussion forums is associated with submitting scientific papers in the field after course completion, but that posting is not associated with submitting papers, even for a MOOC where posting *is* associated with course completion. Assessing not just where a learner is today during a MOOC, but determining how it influences their future career, has the potential to help us better design MOOCs to positively impact students' long-term trajectories.

Assessments in GIFT to Drive MOOC Adaptation

As previously reviewed, assessment techniques applied across MOOCs varies based on the domain being instructed and the activities and exercises configured across lesson interactions. While many of the aforementioned assessments described are aimed at classifying performance states and comprehension levels, it is important to recognize the role these assessments can play in instructional management and remediation practices. With personalization and individualized course-flow serving as a recognized gap in the majority of current MOOC implementations, a current collaborative research effort involving Carnegie Mellon University, the University of Pennsylvania, and the US Army Research Laboratory (ARL) is investigating the utility of the Generalized Intelligent Framework for Tutoring (GIFT) for serving as a framework to structure MOOC content and lessons. In the enhanced MOOC this project is producing, configured assessments across the relevant MOOC-related activities will drive instructional management decisions at the run-time level based on GIFT's pedagogical configurations.

The effort is broken into two phases. The first phase of development focuses on making GIFT LTI compliant, for the purpose of interoperating with large-scale learning management system (LMS) sites like edX. This enables MOOC developers to reference GIFT-managed lessons within the structure and delivery of their course flow, along with the ability to receive data back following the completion of a GIFT lesson for performance tracking and accreditation purposes.

With the LTI component in place, the next phase involves configuring MOOC content into a set of lessons that adhere to the authoring standards and run-time schemas of GIFT. GIFT is unique because it provides a domain-agnostic architecture that enables a course developer to build and sequence content within an instructional design theory that adheres to knowledge development and skill acquisition. GIFT's Engine for Management of Adaptive Pedagogy (EMAP) is a pedagogical model embedded within the 'Adaptive Courseflow' course object, with David Merrill's Component Display Theory (CDT) informing the design (Goldberg, Hoffman & Tarr; 2015; Merrill, 1994). When the EMAP is used, an author configures content for the delivery of "Rules" and "Examples" for each identified concept a lesson targets, followed by configuring two levels of assessment: 1) knowledge recall as it pertains to the declarative and procedural information and 2) skill and application assessment as captured within a set of practice events and/or scenarios.

The "Recall" and "Practice" components of the EMAP can support any number of MOOC related assessment activities, where the derived outcomes of the measures are used to drive what the learner experiences next. The decision involves letting the learner advance to the next configured interaction or remediation logic that is triggered, where the underperforming concepts or states are addressed through an intervention that targets the impasses or misconception identified. The utility of such an approach in a large-scale delivery scenario like an edX MOOC requires experimentation to determine impact and gauge overall effect.

Conclusion

In this chapter, we have briefly discussed the rich state of the art and the relatively more limited state of the practice of assessment in MOOCs. Although some MOOCs today offer calibrated peer assessments, automatically graded essays, step-by-step problem solving, and psychometrically based assessment, most MOOCs continue to base assessment on somewhat arbitrary sets of multiple-choice and fill-in-the-blank items. Furthermore, many of the technologies pioneered in ITSs and other types of artificially intelligent software, such as natural language dialogues, remain unavailable in the MOOC world, and significant barriers exist to bringing them to practice in such systems. Similarly, though there is some work to study metacognition, SRL, and sentiment in MOOCs, MOOC research in these areas has still not reached the level of sophistication seen other areas of educational research.

In some ways, this finding is not surprising. Although there has been an impressive quantity of research conducted on MOOCs, the history of MOOC research remains rather brief. Several of the most impressive demonstrations of the power of assessment in online learning have involved expensive, several-year research efforts. With time, MOOC assessment may reach the same peak in sophistication as ITSs. Having said that, the bigger challenge will be to roll out these benefits to the full diversity of existing MOOCs and use these forms of assessment to drive beneficial intervention. Even in the more mature field of ITSs, it has been challenging to develop interventions that take full advantage of the powerful forms of assessment now available. Solving this challenge in MOOCs will call on the collaboration of both assessment researchers and designers alike.

Acknowledgements

The research described herein has been sponsored by ARL. The statements and opinions expressed in this article do not necessarily reflect the position or the policy of the U.S. Government, and no official endorsement should be inferred.

References

Aleven, V., Mclaren, B., Roll, I. & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education*, *16*(2), 101–128.

Aleven, V., Sewall, J., Popescu, O., Xhakaj, F., Chand, D., Baker, R., Wang, Y., Siemens, G., Rosé, C., Gasevic, D. (2015) The Beginning of a Beautiful Friendship? Intelligent Tutoring Systems and MOOCs. *Proceedings of the 17th International Conference on Artificial Intelligence in Education*, 525–528.

Anderson, J. R., Corbett, A. T., Koedinger, K. R. & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The journal of the learning sciences*, *4*(2), 167–207.

Andres, J.M.L., Baker, R.S., Siemens, G., Gasevic, D., Spann, C.A. (in press) Replicating 21 Findings on Student Success in Online Learning. To appear in *Technology, Instruction, Cognition, and Learning*.

Arnold, K. E. & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267–270). ACM

Azevedo, R. & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia?. *Journal of educational psychology*, *96*(3), 523.

Bachelet, R., Zongo, D. & Bourelle, A. (2015). Does peer grading work? How to implement and improve it? Comparing instructor and peer assessment in MOOC GdP. In *European MOOCs Stakeholders Summit 2015*.

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

Baker, R.S.J.d., Ocumpaugh, J. (2014) Interaction-Based Affect Detection in Educational Software. In R.A. Calvo, S.K. D'Mello, J. Gratch, A. Kappas (Eds.), *The Oxford Handbook of Affective Computing*. Oxford, UK: Oxford University Press.

Bloom, B. S. (1987). A Response to Slavin's Mastery Learning Reconsidered. *Review of Educational Research*, *57*(4), 507–8.

Champaign, J., Colvin, K. F., Liu, A., Fredericks, C., Seaton, D. & Pritchard, D. E. (2014). Correlating skill and improvement in 2 MOOCs with a student's time on tasks. In *Proceedings of the 1st ACM Conference on Learning@Scale* (pp. 11–20). ACM.

Chaplot, D. S., Rhim, E. & Kim, J. (2015). Predicting Student Attrition in MOOCs using Sentiment Analysis and Neural Networks. In *AIED Workshops*.

Chen, G., Davis, D., Lin, J., Hauff, C.,& Houben, G-J. (2016). Beyond the MOOC platform: Gaining Insights aboutLearners from the Social Web. In *Proceedings of the 8th ACM Conference on Web Science*, pp. 15--24, Hannover, Germany. WebSci '16, ACM.

Clarke-Midura, J., Mayrath, M. & Dede, C. (2012). Thinking outside the bubble: virtual performance assessments for measuring inquiry learning.

Cobos, R., Gil, S., Lareo, A., Vargas, F.A. (2016) Open-DLAs: an Open Dashboard for Learning Analytics. *Proceedings of the 3rd Annual Conference on Learning @ Scale*, 265–268.

Colvin, K. F., Champaign, J., Liu, A., Zhou, Q., Fredericks, C. & Pritchard, D. E. (2014). Learning in an introductory physics MOOC: All cohorts learn equally, including an on-campus class. *The International Review of Research in Open and Distributed Learning*, *15*(4).

Corbett, A. T. & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, *4*(4), 253–278.

Dernoncourt, F., Taylor, C., O'Reilly, U. M., Veeramachaneni, K., Wu, S., Do, C. & Halawa, S. (2013). MoocViz: A large scale, open access, collaborative, data analytics platform for MOOCs. In *NIPS Workshop on Data-Driven Education.*

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

Dillon, J., Ambrose, G. A., Wanigasekara, N., Chetlur, M., Dey, P., Sengupta, B. & D'Mello, S. K. (2016). Student affect during learning with a MOOC. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 528–529). ACM.

doignon, J. P. & Falmagne, J. C. (2012). *Knowledge spaces*. Springer Science & Business Media.

Dudycha, A. L. & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology*, *58*(1), 116.

Feng, M., Heffernan, N. & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, *19*(3), 243–266.

Fredericks, C., Lopez, G., Shnayder, V., Rayyan, S. & Seaton, D. (2016, April). Instructor Dashboards In EdX. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 335–336). ACM.

Goldberg, B., Hoffman, M. & Tarr, R. (2015). Authoring Instructional Management Logic in GIFT Using the Engine for Management of Adaptive Pedagogy (EMAP). In R. Sottilare, A. Graesser, X. Hu & K. Brawner (Eds.), *Design Recommendations for Intelligent Tutoring Systems: Authoring Tools (Volume 3)*: US Army Research Laboratory.

Graesser, A. C., Chipman, P., Haynes, B. C. & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, *48*(4), 612–618.

Guo, P. J. & Reinecke, K. (2014). Demographic differences in how students navigate through MOOCs. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 21–30). ACM.

Hood N., Littlejohn A. & Milligan C., (2015). Context Counts: how learners' contexts influence learning in a MOOC, Computers & Education.

Jiang, S., Williams, A., Schenke, K., Warschauer, M. & O'dowd, D. (2014, July). Predicting MOOC performance with week 1 behavior. In *Educational Data Mining 2014*.

Kizilcec, R. F., Pérez-Sanagustín, M. & Maldonado, J. J. (2016). Recommending self-regulated learning strategies does not improve performance in a MOOC. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 101–104). ACM.

Kloft, M., Stiehler, F., Zheng, Z. & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs* (pp. 60–65).

Leelawong, K. & Biswas, G. (2008). Designing learning by teaching agents: The Betty's Brain system. *International Journal of Artificial Intelligence in Education*, *18*(3), 181–208.

Luo, H., Robinson, A. C. & Park, J. Y. (2014). Peer Grading in a MOOC: Reliability, Validity, and Perceived Effects. *Journal of Asynchronous Learning Networks, 18*(2).

Merrill, M. D. (1994). *The descriptive component display theory*: Educational Technology Publications, Englewood Cliffs, NJ.

Mislevy, R. J. & Riconscente, M. M. (2006). Evidence-centered assessment design. *Handbook of test development*, 61–90.

Mitros, P. , et. al. (2013). Teaching Electronic Circuits Online: Lessons from MITx's 6.002x on edX. IEEE ISCAS.

Mitros, P., Agarwal, A., Paruchuri, V. (2014) Ubiquity symposium: MOOCs and technology to advance learning and learning research: assessment in digital at-scale learning environments. *Ubiquity*, 1–9.

Onah, Daniel F. O. and Sinclair, Jane (2016) Exploring learners' strategies of self-regulated learning abilities in a novel MOOC Platform : eLDa. In: 23rd Annual Conference of the Association for Learning Technology (ALT2016), University of Warwick, United Kingdom, 6–8 Sep 2016.

Pardos, Z., Bergner, Y., Seaton, D. & Pritchard, D. (2013, July). Adapting Bayesian knowledge tracing to a massive open online course in edx. In *Educational Data Mining 2013*.

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

Roll, I., Aleven, V., McLaren, B. M. & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction, 21*(2), 267–280.

Salden, R. J., Koedinger, K. R., Renkl, A., Aleven, V. & McLaren, B. M. (2010). Accounting for beneficial effects of worked examples in tutored problem solving. *Educational Psychology Review*, *22*(4), 379–392.

Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P. & Pritchard, D. E. (2014). Who does what in a massive open online course?. *Communications of the ACM*, *57*(4), 58–65.

*S*harkey, M. & Sanders, R. (2014, October). A process for predicting MOOC attrition. *In Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs* (pp. 50–54).

Shih, B., Koedinger, K. R. & Scheines, R. (2011). A response time model for bottom-out hints as worked examples. *Handbook of educational data mining*, 201-212.

Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S.B., Ferguson, R., Duval, E., Verbert, K., Baker, R.S.J.d. (2011) Open Learning Analytics: an integrated & modularized platform: Proposal to design, implement and evaluate an open platform to integrate heterogeneous learning analytics techniques. Athabasca, Alberta, Canada: Society for Learning Analytics Research.

Viswanathan, R. (2012). Teaching and Learning through MOOC. *Frontiers of Language and Teaching, 3*(1), 32-40.

Wang, Y.E., Baker, R., Paquette, L. (2017) Behavioral Predictors of MOOC Post-Course Development. *Proceedings of the Workshop on Integrated Learning Analytics of MOOC Post-Course Development*.

Wang, Y. & Heffernan, N. (2013). Extending knowledge tracing to allow partial credit: Using continuous versus binary nodes. In *International Conference on Artificial Intelligence in Education* (pp. 181-188). Springer Berlin Heidelberg.

Wen, M., Yang, D. & Rose, C. (2014). Sentiment Analysis in MOOC Discussion Forums: What does it tell us? *Proceedings of the International Conference on Educational Data Mining*.

Winne, P. H. & Baker, R. S. (2013). The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *JEDM-Journal of Educational Data Mining*, *5*(1), 1-8.

Yang, D., Sinha, T., Adamson, D. & Rosé, C. P. (2013, December). Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop* (Vol. 11, p. 14).