This is a post-print of the document published in:

Alexandron, G., Wiltrout, M. E., Berg, A., & Ruipérez-Valiente, J. A. (2020). Assessment that matters: balancing reliability and learner-centered pedagogy in MOOC assessment. In LAK '20: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (pp. 512–517).

DOI: https://doi.org/10.1145/3375462.3375464

https://dl.acm.org/doi/abs/10.1145/3375462.3375464

© 2020 ACM

Assessment that matters: Balancing reliability and learner-centered pedagogy in MOOC assessment

Giora Alexandron*
Weizmann Institute of Science
giora.alexandron@weizmann.ac.il

Mary Ellen Wiltrout*
Massachusetts Institute of Technology
mew27@mit.edu

Aviram Berg
Weizmann Institute of Science
aviram.berg@weizmann.ac.il

José A. Ruipérez-Valiente
University of Murcia
jruiperez@um.es

ABSTRACT

Learner-centered pedagogy highlights active learning and formative feedback. Instructors often incentivize learners to engage in such formative assessment activities by crediting their completion and score in the final grade, a pedagogical practice that is very relevant to MOOCs as well. However, previous studies have shown that too many MOOC learners exploit the anonymity to abuse the formative feedback, which is critical in the learning process, to earn points without effort. Unfortunately, limiting feedback and access to decrease cheating is counter-pedagogic and reduces the openness of MOOCs. We aimed to identify and analyze a MOOC assessment strategy that balances this tension between learner-centered pedagogy, incentive design, and reliability of the assessment. In this study, we evaluated an assessment model that MITx Biology introduced in a MOOC to reduce cheating with respect to its effect on two aspects of learner behavior – the amount of cheating and learners' engagement in formative course activities. The contribution of the paper is twofold. First, this work provides MOOC designers with an 'analytically-verified' MOOC assessment model to reduce cheating without compromising learner engagement in formative assessments. Second, this study provides a learning analytics methodology to approximate the effect of such an intervention.

CCS CONCEPTS

• **Applied computing** → **E-learning**; • **Unsupervised learning** → *Anomaly detection.*

KEYWORDS

MOOCs, Assessment, Learning Analytics

*Both authors contributed equally to the paper

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
LAK '20, March 23–27, 2020, Frankfurt, Germany
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7712-6/20/03…$15.00
https://doi.org/10.1145/3375462.3375464

ACM Reference Format:
Giora Alexandron, Mary Ellen Wiltrout, Aviram Berg, and José A. Ruipérez-Valiente. 2020. Assessment that matters: Balancing reliability and learner-centered pedagogy in MOOC assessment. Submitted to *LAK'20* . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3375462.3375464

1 INTRODUCTION

Learner-centered pedagogy highlights formative assessment that includes personalized and immediate feedback that can help learners self-regulate their learning [20]. It is widely accepted among educators that using external incentives is an effective way to motivate learners to engage in pedagogically-desired activities (a simple example is grading homework to motivate individuals to practice so that they learn during the semester, rather than cramming for the final exam). The need for learner self-regulation and external motivators is especially applicable to MOOCs, since these attract a heterogeneous learner population but learners typically receive little individual support. Even if learners have genuine interest in learning ('real learners') [4], they might have strategic goals such as earning a certificate that can affect their learning behavior. For novices, providing the correct incentives is especially important, as they need to develop their self-regulated learning skills [9, 14].

While MOOCs provide access to quality learning materials for everyone with an internet connection, the platform companies and institutions providing the courses are seeking financial sustainability to continue to make the resource available for free or at a low cost. Currently, MOOC certification is the cornerstone of the financial model [26]. Thus, monetization depends on the perception of certificates as a reliable and valid proof of proficiency [2].

However, previous studies indicated that in MOOCs that follow a learner-centered design, a significant amount of learners exploit learning-centered features, such as graded formative assessment with feedback, to achieve a certificate with less effort [24]. Preventing such cheating is typically done by limiting the feedback or access to the formative assessment. This means that design for reliability can require sacrificing design best for learning.

Thus, **the main goal of our research is to develop and validate MOOC assessment models that support learner-centered pedagogy, without sacrificing trustworthiness of the assessment for certification**. The current research studies such a model, the Competency Exam (CE), introduced by MITx 7.00x "Introduction to Biology" (abbreviated as "MITx 7.00x" going forward).

1.1 The Competency Exam Assessment Model

After several course runs using the traditional MOOC structure in MITx 7.00x, and motivated to maintain academic rigor while continuing to share knowledge with others for free, MITx Biology introduced the CE assessment model in July 2016 [22]. The CE model separates learning from certification and combines many assessment design techniques that render cheating more difficult. While the original course materials (videos, concept questions, problem sets, exams) remained available for self-paced learning for free, learners had to pay for verified track to access the CE available during a specific week. Additional assessment design aimed to reduce opportunity for cheating included setting the CE as a timed exam, closing the discussion forum during the CE week, hiding feedback on correctness, eliminating access of the CE post-completion, and randomizing questions at the page level across 20 different pages of multi-part questions that tested learning objectives from across course topics. More details on the CE and MITx 7.00x are provided in Section 2.

Research Rationale. Besides the expected and desired reduction of cheating, the CE model may also drive undesired changes in learners' behavior, as a consequence of the effect of the new grading model on learners' incentives. First, we expected that making the within-course formative assessment ungraded would potentially reduce learners' motivation to spend time and effort attempting the problems. Second, we expected that making the learning materials self-paced and ungraded may cause some learners to cram their learning towards the date of the CE release. Third, we expected that the effect of less engagement with the formative assessment items and cramming could lead to less learning, and as a result, to a decrease in the certification rate.

1.2 Research Questions

Guided by the research rationale defined, our research seeks to answer the following research questions (RQs) regarding the impact of the CE assessment model on the following factors, relative to the previous assessment model:

- RQ1: Does the CE model reduce the amount of cheating?
- RQ2: Does the CE model reduce the level of learner engagement with within-course formative assessment items?
- RQ3: Does the CE model affect the way learners spread their learning over time?
- RQ4: Does the CE model reduce the certification rate?

1.3 Related Work

With online learning in general and MOOCs in particular gaining a lot of influence within the educational ecosystem, numerous studies have addressed the issue of implementing assessment at scale with low intervention from experts [1]. A review of the native assessment functionalities in edX, Coursera, MiriadaX and Khan Academy revealed that most activities consist of very simple auto-graded items [19]. As it is difficult to assess higher-order cognitive skills with such items, designing MOOCs that stimulate meaningful learning is challenging, though feasible [8]. In addition, MOOCs that follow a problem-centric, active learning approach are rated as more engaging [12]. Thus, researchers and practitioners are implementing and testing more learner-centered assessment alternatives beyond the conventional ones [10]. A common approach is combining formative assessment with more traditional, content-oriented summative assessment for certification [7]. However, if these formative assessment activities are completely optional, the learner participation decreased significantly in one study [15]. And if the formative assessments are graded to incentivize active learning, some learners will abuse the course assessment model to earn credit without effort [2], relying on their anonymity [16]. This raises the tension between formative assessment that prioritizes learning and summative assessment that prioritizes accountability [6].

Indeed, there is a growing body of research on cheating in MOOCs, using methods such as plagiarism [11], multiple accounts [2, 21, 24] or passing responses among peers [25]. The application of these methods suffer from the common limitation that their design detects certain types of cheating while failing to detect different ones. To address this, a recent study suggested a method that is based on anomaly detection [3]. The idea behind anomaly-based method is that cheating leaves a statistical signature that manifests as aberrant behavior, which one can capture using person-fit statistics [5, 17]. Because such a strategy is indirect, anomaly detection should be used with great caution when applied to an individual, but as shown in [13], the risk is decreased, and the approach can be very effective when used as an aggregated measure of group behavior. Our study implements the method suggested in [3] but applies the technique to perform an aggregated estimation of cheating.

With MOOCs pivoting towards professional development courses and for-credit online degrees [23], the necessity to develop learner-centered assessment methods that are robust against cheating is stronger than ever. Previous work tackled the effect of assessment in pedagogy and its relationship with cheating. However, this is the first MOOC paper that actually measures the effect of assessment design on cheating and discusses the tension between assessment, pedagogy, and monetization in MOOCs.

2 BACKGROUND - THE COURSE AND ASSESSMENT MODELS

Course structure and early assessment model. The 7.00x "Introduction to Biology" MOOC [1] first released during March 2013. The course structure followed a model similar to the on-campus counterpart taken by enrolled MIT students. The triweekly lectures translated online into a sequence of short videos interspersed with concept questions referred to as test yourself questions. The discussion forum and deep dive videos substituted for teaching assistant-led, working sessions. The problem sets gave learners the chance to apply the concepts from the videos, while the exams assessed the learners' skills in the subject. Like the on-campus equivalents, learning rather than certification (or grading) was central to the purpose of the videos, test yourself questions, discussion forum, and problem sets. The summative exams counted for the majority (80%) of the overall grade to earn a certificate, but the formative problem sets counted toward the grade (20%) as well. To pass the course for a certificate, a learner had to earn an overall score of 60%. Both problem sets and exams had limited attempts, gave immediate correctness feedback, and provided show answer with elaborated feedback after a deadline.

[1] https://www.edx.org/course/introduction-to-biology-the-secret-of-life-4

Tension between learning and cheating. From the start, a tension between setting practices best for learning versus those best for preventing cheating existed. For example, the course team noted the ease of multiple-account cheating [21, 24], and therefore delayed or restricted some of the formative feedback to limit the temptation for learners to try the strategy to increase scores. With such practices, some important advantages of online learning are taken away, as learners could only access answers and explanations or discuss an answer on a forum after that assignment's due date passed, and not in the moment of struggle with a concept.

At the same time, this design was not very robust against cheating. Unlike on campus, the MOOC version of the course included the same exams each new run with minor revisions for clarity or technical updates. So even if answers became available after a deadline, some learners had access to answers from previous runs of the course. Therefore, MITx Biology felt the need to separate the learning components of the course from the assessment of a learner to award a certificate to maintain rigor in the course and integrity of those earning certificates.

The Competency Exam rationale. Not bound to the schedule or policies of the on-campus course and motivated by the edX policy change to charge for certificate-track registration, MITx Biology began the CE model [2] for running the MITx 7.00x MOOC in the MIT spirit to continue to share knowledge with the world for free. Under this model, learners could work through all existing course content as ungraded in a self-paced manner with instant feedback on problem correctness as well as more detailed explanations after trying a question at least once. Only if a learner wanted to try to earn a certificate, then the learner could pay the fee to take the graded CE offered during a specific week about every four months.

The Competency Exam features. The timed CE restricts access to learners in the paid verified track, has no public discussion forum support (email only), gives no feedback on correctness, provides no access after completion, includes randomization of entire story-based problems, and ties together learning objectives across the course while balancing the coverage. Even for the biology, the CE model eliminates the artificial restriction to ask questions only on blocks of content that fit in a time window prior to an exam versus how scientists may think about the topics taught across the course in an integrated manner. To pass the CE for a certificate, a learner has to earn an overall score of 80%.

3 METHODOLOGY

To evaluate the effect of the CE assessment model on the behaviors of learners (see Research Questions, Subsection 1.2), we compared the relevant metrics on course runs before and after introduction of the CE model. In the following sections, we describe the data and the learning analytics methods.

3.1 Data

Course runs. The context of this research is the MITx 7.00x MOOC (see Section 2). We used data from five consecutive runs of the course, and we refer to them as noCE, CE1, CE2, CE3, and CE4. The noCE run is the last run *before* the introduction of the CE assessment model, and CE1, CE2, CE3, and CE4 are the first four runs of the course *with* the CE model. We did not include runs prior to 2014 due to technical issues and availability of data or runs later than CE4 since four runs seemed sufficient in the study.

Research Population. This study focuses on *verified* learners. Verified learners are those who registered for the paid track and therefore have the opportunity to earn a certificate when earning a passing score in a course. The number of verified learners in each run is 296 (noCE), 35 (CE1), 83 (CE2), 82 (CE3), 98 (CE4). For the analyses, we focus on the *certified* learners among the verified registrants when referring to the amount of cheating and amount of items done. During the noCE run, edX still offered the 'honor' certificate, however, to have a comparable group of learners between runs, we did not include these learners.

Data Mining. This study uses behavioral statistics that are computed from learners' clickstream data, which mainly include video events (play, pause, etc.), responses to assessment items, and navigation to course pages. In addition, we use the course structure files that hold information that describes the elements of the course.

3.2 Methods

To evaluate the amount of cheating, we used the anomaly detection method as proposed and evaluated in [3]. The method is based on person-fit statistics that measure learners' aberrant behavior. Its theoretical foundation relies on the psychometric literature (see Section 1.3). In brief, the idea is that cheating leaves a statistical signature in the data because the learners' probability of answering an item correctly, or the time they spend on that item, given its level of difficulty, differ between cheated-upon items and non-cheated items. This is captured using two measures: Guttman error and time-on-task Guttman error.

Guttman Error (Gnormed). Guttman error is defined as the number of item pairs in which a learner answers an easier item incorrectly and a more difficult item correctly, normalized by the total number of pairs [17]. We used a non-parametric model implemented in *R*'s *PerFit* [27]. It is computed on all possible pairs.

Time-on-Task Guttman Error (Gnormed-time). This is a time-based person-fit statistic that applies the idea of Guttman Error to time-on-task. *Gnormed-time* approximates, per user u, the number of item pairs i,j in which the time ordering of u on i,j is opposite to the average time ordering of i,j, normalized by the number of pairs.

The procedure for computing *Gnormed-time* is taken from [3]. This method is based on a reduction of the time-on-task matrix into a Boolean matrix M, where $M[u, i]$=0 if the time-on-task of u on i is larger than the mean time-on-task of u, and 1 otherwise. This means that an item on which many learners spent more time than usual will have a lot of 0s in that column, the equivalence to a 'hard' item. We then computed *Gnormed-time* by applying the Guttman-error algorithm on M.

Gnormed-time provides a simple and efficient alternative to more sophisticated response time models such as the log-normal [28], which are difficult to fit on MOOC data.

A previous study [3] demonstrated that high *Gnormed* and *Gnormed-time* strongly correlated with cheating, and thus, their relative levels provide indirect measures of the amount of cheating in the course.

[2] https://web.mit.edu/mitxbio/700CEx.html

4 RESULTS AND ANALYSIS

In this section, we present the results of analysis aimed at addressing the research questions defined in Subsection 1.2.

4.1 RQ1: Amount of Cheating

We evaluated the change in the amount of cheating by measuring the change in two aberrant behavior person-fit statistics: *Gnormed* and *Gnormed-time*, as described in Section 3.

Figure 1 shows the distribution of *Gnormed* on the left and *Gnormed-time* on the right among certified learners in the five runs of the MITx 7.00x MOOC between 2014 and 2017. We conducted a one-way ANOVA with linear contrasts to compare the effect of the CE on both *Gnormed* and *Gnormed-time*. We carried out a Levene's test, and the results did not meet the homogeneity of variance assumption, hence we conducted the analysis accordingly. The analysis showed that the effect of the CE on *Gnormed* and *Gnormed-time* was statistically significant in both cases ($F(4, 217) = 43.981$, $p = .000$ and $F(4, 217) = 5.833$, $p = .000$ respectively).

This confirms that there was a sharp decrease in both *Gnormed* and *Gnormed-time* statistics after the introduction of the CE assessment model, relative to the original course assessment model.

4.2 RQ2: Engagement with Within-course Items

In the course with the original assessment model, the formative assessments counted toward the final grade with the exams, while in the runs that use the CE assessment model, only the CE score counts for the final grade. Since the formative assessments are ungraded in the CE assessment model, we first hypothesized that not crediting these items for the final grade will decrease learners' engagement with these practice problems.

Figure 2 displays the fraction of formative assessment items attempted by certified learners in each of the runs. We conducted a one-way ANOVA with linear contrasts to compare the effect of the CE model on the fraction of formative assessment items that certified learners attempted. The analysis showed no significant difference ($F(4, 217) = 0.992$, $p = .413$). Thus, we accepted the null hypothesis that there is no difference in the fraction of items attempted by learners before and after the introduction of the CE model. We further discuss this in Subsection 4.5.

4.3 RQ3: Activity Level Over Time

We also examined when certified learners interacted with the course materials. Without deadlines spaced over time for all of the formative assessments in the CE model, we hypothesized that learners follow a less regular pattern of course completion over time. Figure 3 demonstrates how learners spread their learning over time, by focusing on the number of daily events, normalized by the number of learners.

As shown, learning activity over time changed significantly in the CE course runs versus noCE run. While in the noCE run, learners' activity spread regularly over the entire course, introduction of the CE (runs CE1, CE2, CE3, and CE4) corresponded to relatively low learners' activity during most of the course run period, with a cramming pattern (very high activity in a short period of time) towards the week of the CE.

4.4 RQ4: Certification Rate

Last, we tested the hypothesis that the CE assessment model reduced the certification rate compared to the original course assessment model. With a higher score threshold to pass (80% versus 60%) and the CE testing objectives throughout all of the course topics, we expected that learners pass at a lower rate. Figure 4 presents the fraction of verified learners who met the criteria to earn a certificate. For example, in noCE, 143 out of 296 verified users (48%) earned a certificate.

As shown, the fraction of certified learners dropped significantly, from ∼ 48% before the CE assessment model, to ∼ 26% in the first four runs after the introduction of the CE assessment model in support of our initial hypothesis.

4.5 Discussion

Regarding the first research question, the findings presented in Subsection 4.1 provide strong evidence that the CE assessment model reduced the amount of cheating in the MITx 7.00x MOOC. The fact that the CE is effective against cheating is not surprising, as the assessment model combines several means that render cheating very difficult, such as randomization (described in Section 2). However, using robust learning analytics methods to verify that the design achieves its goals is the cornerstone of evidence-based design.

Our second research question aims to study the effect of the CE assessment model on learners' motivation to engage with the in-course formative assessment, now ungraded. The findings presented in Subsection 4.2 show no effect of the CE on the fraction of formative assessment items that certified learners attempted. We interpret this as an evidence that the CE did not decrease learners' motivation to use formative assessment. This finding is somewhat surprising, because we expected that learners would be less motivated to invest time and effort in practicing their knowledge if it did not contribute to their grade. This finding might indicate that the verified users (our research population) are serious learners who follow the course design.

However, on the negative side, referring to the third research question, our conclusion is that the CE does lead learners to cram their learning, as shown in Subsection 4.2. It is well-documented in educational research that cramming learning to a shorter time window is less effective than spacing learning over time, and this was also shown in MOOCs (e.g., [18]).

Learners tendency to cram their learning towards the CE might explain some of the reduction in the certification rate among the verified users – the fourth research question – shown in Subsection 4.4. However, the fact that in the CE model learners need 80% to pass versus 60% in the original course model may also explain the decrease in certification rate.

4.6 Limitations

With respect to the reduction in cheating, the main limitation is that, without having direct measures, evaluating the amount of cheating relies on indirect methods. However, reviewed in Subsection 1.3, this approach is rooted in psychometrics and MOOC literature. Also, since the change in the aberrant behavior measures is very large, we believe that the effect is significant as well.

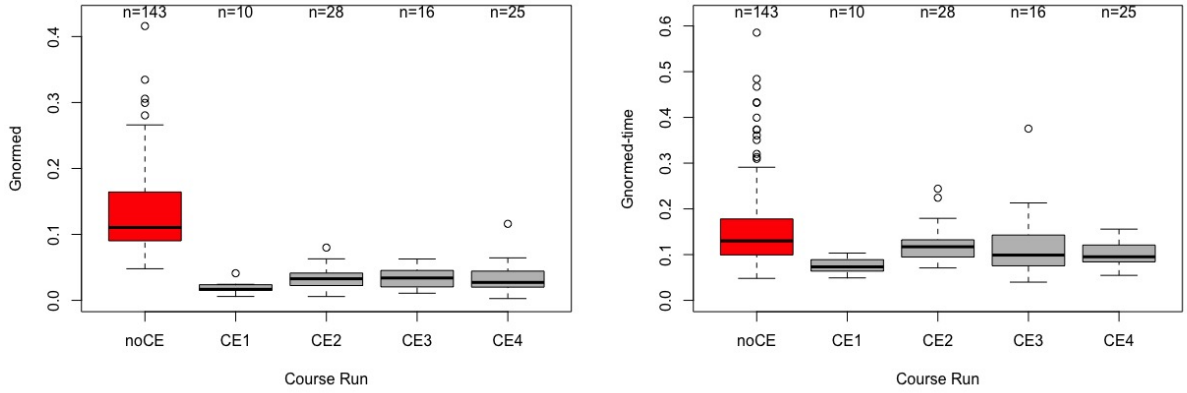

Figure 1: Distribution of *Gnormed* (left) and *Gnormed-time* (right) among verified certificate earners (*n* denotes their number per run). The red bar presents the data for the noCE run, and the gray bars indicate the results for the CE runs.

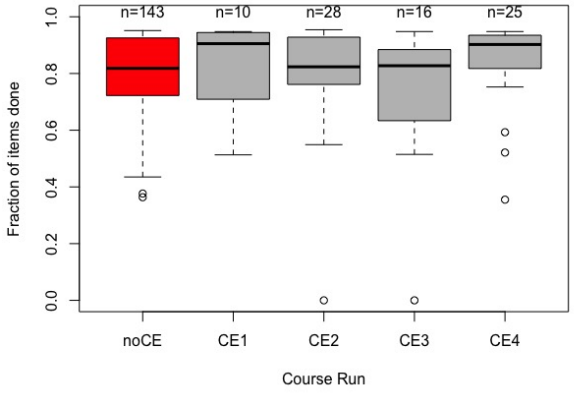

Figure 2: The fraction of formative assessment items attempted by verified certificate earners.

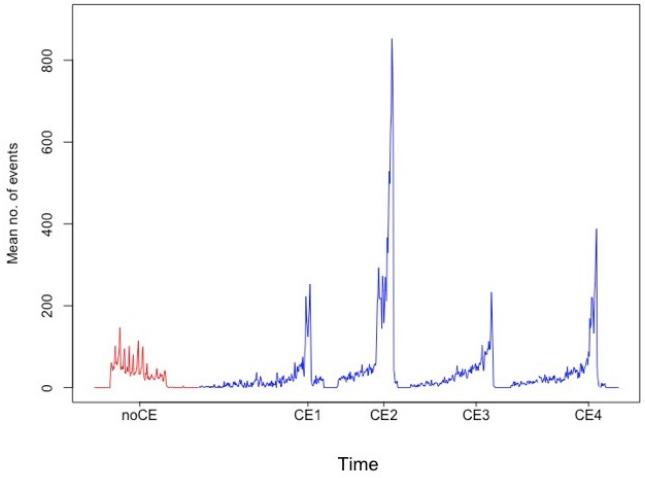

Figure 3: Amount of daily activity, by course run.

4.7 Future Research

In future research, we intend to focus on several important aspects. One is examining the difference between the certified learners and the non-certified ones. Since the non-certified verified learners paid for the chance to earn a certificate – a strong indication of their motivation to succeed in the course – but did not qualify for one by final grade, it would be interesting to understand the barriers for certification among them.

Our next step is to evaluate a design with mixed incentives, for example, a design that combines both a CE assessment model and graded formative assessment, with both components as 'necessary but insufficient' condition for earning a certificate.

5 SUMMARY AND CONCLUSIONS

Assessment in MOOCs has an inherent tension between pedagogy and trustworthiness. Orienting towards 'pedagogy' emphasizes formative assessment that is graded to incentivize active participation and instructional feedback to support learning. However, since research shows that learners can abuse graded formative assessment in MOOCs, maintaining trustworthy assessment for certification pushes towards limited access and no feedback.

Our goal is studying MOOC assessment models that support and motivate learning without compromising reliability for certification. Within this overarching goal, this paper studies an assessment

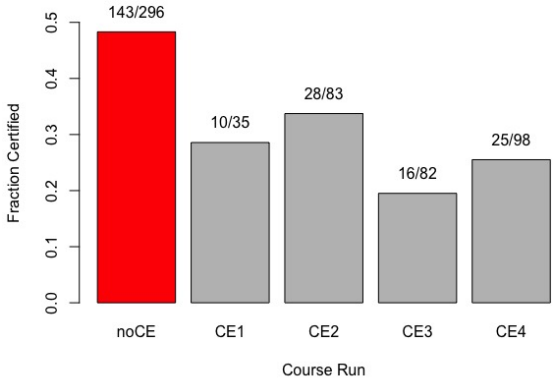

Figure 4: Fraction of verified learners who earned a certificate.

model – The Competency Exam (CE), introduced by MITx Biology in 2016 – that balances secured-examination and formative assessment.

We study the effect of the CE on the amount of cheating and on learners engagement. Our findings demonstrate that i) the CE significantly decreases the amount of cheating; ii) (certified) learners engagement with the formative assessment items remain the same, even though these items did not count for the final grade; iii) learners tended to cram their learning into a short period before the CE; and iv) the percentage of verified learners who met the certification criteria dropped from $\sim$ 48% before the CE, to $\sim$ 26%.

For MOOC designers, we recommend adopting the CE assessment model if the goals are providing access to formative assessment for learning, while maintaining rigorous assessment for certification. However, our results suggest that to incentivize learners to spread their learning rather than to cram for the final, it would be good to include the formative assessment in the overall course grade, with both the CE and the formative assessment as 'necessary but insufficient' conditions for certification.

In addition to providing an evaluation of a novel MOOC assessment model, this study also presents a learning analytics methodology to approximate the trustworthiness of such models.

ACKNOWLEDGMENTS

The authors thank Sa'ar James Karp Gershon for statistical consultation. The Israeli Ministry of Science and Technology supported the research of GA and AB under project no. 3-15607.

REFERENCES

[1] Wilfried Admiraal, Bart Huisman, and Olga Pilli. 2015. Assessment in Massive Open Online Courses. *Electronic Journal of E-learning* 13, 4 (2015), 207–216.

[2] Giora Alexandron, José A. Ruipérez-Valiente, Zhongzhou Chen, Pedro J. Muñoz-Merino, and David E. Pritchard. 2017. Copying@Scale: Using Harvesting Accounts for Collecting Correct Answers in a MOOC. *Computers and Education* 108 (2017), 96–114.

[3] Giora Alexandron, José A. Ruipérez-Valiente, and David E. Pritchard. 2019. Towards a General Purpose Anomaly Detection Method to Identify Cheaters in Massive Open Online Courses. In *Proceedings of the 12th International Conference on Educational Data Mining*. 480–483.

[4] Giora Alexandron, Lisa Y. Yoo, José A. Ruipérez-Valiente, Sunbok Lee, and David E. Pritchard. 2019. Are MOOC Learning Analytics Results Trustworthy? With Fake Learners, They Might Not Be! *International Journal of Artificial Intelligence in Education* (2019).

[5] Dmitry I. Belov. 2016. Comparing the Performance of Eight Item Preknowledge Detection Statistics. *Applied Psychological Measurement* 40, 2 (2016), 83–97.

[6] Paul Black, Christine Harrison, Clare Lee, Bethan Marshall, and Dylan Wiliam. 2004. Working inside the Black Box: Assessment for Learning in the Classroom. *Phi Delta Kappan* 86, 1 (2004), 8–21.

[7] David Carless. 2015. Exploring learning-oriented assessment processes. *Higher Education* 69, 6 (2015), 963–976.

[8] Kimberly Colvin, John Champaign, Alwina Liu, Qian Zhou, Colin Fredericks, and David Pritchard. 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 (2014).

[9] Yossi Ben David, Avi Segal, and Kobi Gal. 2016. Sequencing Educational Content in Classrooms Using Bayesian Knowledge Tracing. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16)*. 354–363.

[10] María del Mar Sánchez-Vera and María Paz Prendes-Espinosa. 2015. Beyond objective testing and peer assessment: alternative ways of assessment in MOOCs. *International Journal of Educational Technology in Higher Education* 12, 1 (2015), 119–130.

[11] L Gibbs. 2012. Yes, Plagiarism: How Sad is That. *Coursera Fantasy: Blogging My Way through a MOOC* (2012).

[12] Khe Foon Hew. 2016. Promoting engagement in online courses: What strategies can we learn from three highly rated MOOCS. *British Journal of Educational Technology* 47, 2 (2016), 320–341.

[13] Brian A Jacob and Steven D Levitt. 2003. *Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating*. Working Paper 9413. National Bureau of Economic Research.

[14] Kim Kelly and Neil Heffernan. 2015. Developing Self-regulated Learners Through an Intelligent Tutoring System. In *Artificial Intelligence in Education*, Cristina Conati, Neil Heffernan, Antonija Mitrovic, and M. Felisa Verdejo (Eds.). Springer International Publishing, 840–843.

[15] Jonathan Kibble. 2007. Use of unsupervised online quizzes as formative assessment in a medical physiology course: effects of incentives on student participation and performance. *Advances in Physiology Education* 31, 3 (2007), 253–260.

[16] Chula G King, Roger W Guyette Jr, and Chris Piotrowski. 2009. Online exams and cheating: An empirical analysis of business students' views. *J. of Educators Online* (2009).

[17] Rob R. Meijer. 1994. The Number of Guttman Errors as a Simple and Powerful Person-Fit Statistic. *Applied Psychological Measurement* 18, 4 (1994).

[18] Y. Miyamoto, C. Coleman, J. Williams, J. Whitehill, S. Nesterko, and J. Reich. 2015. Beyond Time-on-Task: The Relationship between Spaced Study and Certification in MOOCs. *Journal of Learning Analytics* (2015).

[19] Pedro J. Muñoz-Merino, José A. Ruipérez-Valiente, Juan Luis Sanz Moreno, and Carlos Delgado Kloos. 2015. Assessment activities in massive open on-line courses: Assessment activities in MOOCs. In *Furthering higher education possibilities through massive open online courses*. IGI Global, 165–192.

[20] David J. Nicol and Debra Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education* 31, 2 (2006), 199–218.

[21] Curtis G Northcutt, Andrew D Ho, and Isaac L Chuang. 2016. Detecting and preventing "Multiple-Account" cheating in massive open online courses. *Computers & Education* 100 (2016), 71–80.

[22] Office of Digital Learning. 2017. Competency exam certificate now available for MITx Introduction to Biology course. http://news.mit.edu/2016/competency-exam-available-for-mitx-introduction-to-biology-0103. Accessed: 2019-08-22.

[23] Justin Reich and José A Ruipérez-Valiente. 2019. The MOOC pivot. *Science* 363, 6423 (2019), 130–131.

[24] José A. Ruipérez-Valiente, Giora Alexandron, Zhongzhou Chen, and David E. Pritchard. 2016. Using Multiple Accounts for Harvesting Solutions in MOOCs. *Proceedings of the Third ACM Conference on Learning @ Scale - L@S '16* (2016), 63–70.

[25] José A Ruipérez-Valiente, Srecko Joksimović, Vitomir Kovanović, Dragan Gašević, Pedro J Muñoz-Merino, and Carlos Delgado Kloos. 2017. A data-driven method for the detection of close submitters in online learning environments. In *Proceedings of WWW'17 Companion*. 361–368.

[26] Dhawal Shah. 2019. Coursera's Monetization Journey: From 0 to $100+ Million in Revenue. https://www.classcentral.com/report/coursera-monetization-revenues/. Accessed: 2019-09-01.

[27] Jorge N. Tendeiro, Rob R. Meijer, and A. Susan M. Niessen. 2016. PerFit: An R Package for Person-Fit Analysis in IRT. *Journal of Statistical Software* 74, 5 (2016), 1–27. https://doi.org/10.18637/jss.v074.i05

[28] Wim J. van der Linden. 2006. A Lognormal Model for Response Times on Test Items. *Journal of Educational and Behavioral Statistics* 31, 2 (2006), 181–204.