The New Educational Review

Enrique Sánchez Acosta,
Juan José Escribano Otero,
Gabriela Christie Toletti
Spain

Peer Review Experiences for MOOC. Development and Testing of a Peer Review System for a Massive Online Course

Abstract

Although at first MOOC (Massive Open Online Courses) did not use peer reviews, this kind of assessment has increasingly demonstrated the benefits that it can contribute to this type of course by improving the learning process, increasing decisions making abilities, and developing several other academic skills. Other MOOC assessment instruments do not provide students with these opportunities. This paper discusses the results obtained by the most commonly used massive online course platforms, detailing their features and limitations, as well as the experience in the implementation and use of a peer review system for a course of more than 7300 students. This study also comments on how evaluation rubrics are created, along with the final results, and the impact of the inclusion of this type of evaluation in MOOC.

Keywords: *MOOC, evaluation, peer, automatism, massive*

It is first necessary to define and frame the concept of peer review, currently used by most scientific journals in the context of massive online courses. The evaluation system of scientific work by community members called peer review or referee system is a process that begins when a scientist submits an article to a magazine editor with the intent of it being published. Then selected specialists (referees) evaluate the quality of the work and determine if the product of research has potential for the stated purpose, or if some additional work has to be done before publication. (Mestaza and Cuevas, 2002)

However, in massive online courses such types of assessment have been distorted. It can be seen from the above definition given by Cuevas and Mestaza, how the word "specialist" is specified; however, in online courses students themselves are often the ones who try to evaluate their peers. To demonstrate that these evaluations are equally valid as if they were conducted by a specialist, current MOOC supported platforms are based on the large number of evaluations of an exercise that these students can perform to determine a more accurate rating. Thus, virtually all platforms support these types of assessment, and they all allow for increasing the number of times some work is evaluated to a number superior to two.

It is difficult to frame peer reviews within the assessment instruments used in MOOC, therefore in order to better define the peer evaluation process, a division of assessment instruments into three basic types is proposed:

Automation based tools

These tools or assessment instruments are based on automatic programs that analyze the responses with tools that implement a default correction algorithm. With these tools, reliability of correction is pursued so that the same answer will receive the same evaluation every time it is subjected to automation. There are different types of instruments that can fit in this category, but the key feature is that they do not require human intervention, making them particularly suitable to be used in MOOC. Examples might be: multiple choice tests, automatic evaluation of problem sets, programming tasks, surveys and questionnaires, attitudes rating scales, written exams, troubleshooting, comparison charts, and images. In free writing responses, semantic analyzers can be used with or without dictionaries and thesauruses.

MIT (Massachusetts Institute of Technology) is conducting research (for their Edx platform) on various Text Analysis Systems or AEG (Automated Essay Grading) (Markoff, 2013) to allow for essays and written tests to be also automatically evaluated.

This approach, of course, also has plenty of detractors like those grouped within HumanReaders.org. This group has already gathered more than 4,000 signatures of professionals from different universities around the world. They are carrying out a call to all schools and universities to stop using automatic correction tools for written work, especially in the case of written exams or tests that are critical for student graduation. Their main argument is that computers cannot read and cannot measure the essential elements of written communication such as:

accuracy, reasoning, matching evidence, common sense, ethical stance, deciding if an argument is compelling, organization of concepts, clarity, and accuracy, among other things. ("Human Readers," 2013)

However, there are several of these types of systems currently on the market and we should not forget that machines are much more consistent and can evaluate a larger number of items in a shorter period of time (Ezeiza, A, 2013). Currently, these systems combine algorithmic methods of grammatical analysis with sematic analysis, and holistic methods based on word searches. For example, the Summary Street System (Steinhart, 2000) compares summaries with the original text, or the Computer Learner Corpora (Granger, Hung, and Petch-Tyson, 2002) compiles a database of students' texts to compare and analyze other written work. The e-rater (Attali and Burstein, 2006) combines statistical analysis and natural language processing to contrast the results with its database; it examines grammar issues, discourse markers, and lexical content using about 100 indicators. The results are supposed to have a success rate between 84% and 94% compared to human evaluators. This system is driven by ETS (Educational Testing Service) to develop the Criterion program. ETS uses this system in well-known TOEFL tests (Test of English as a Foreign Language), matching machine with human evaluator only for some specific tests, which saves a significant amount of money (Knoch, 2009).

Tools based on authority

These are the tools which involve a professional or a person skilled in the field. They are very difficult to implement in a MOOC, mainly due to the large number of students enrolled in the course, so this type of evaluation would require an enormous amount of time from a professor or professors. However, sometimes these corrections are delegated to dynamic adjunct instructors who energize and support students. The problem of evaluation criteria disparity appears when a large group of professors is in charge of correcting instead of just one professor, this could make the same response receive very different evaluations depending on the faculty member evaluating and even depending on when the faculty member performs the evaluation. To alleviate this problem, it is possible to apply very sophisticated evaluation rubrics that determine more objective corrections, parameters, and descriptors. But in the end, human beings evaluate largely based on intuition. Some authors argue that evaluators' previous experience and knowledge are more valuable and relevant than any descriptor or rubric. Therefore, rather than spending hours and studies to build reliable and valid rubrics, they believe that it is more profitable to spend that money and effort on preparing people who can evaluate tests, reach a degree of agreement, and handle scales (Ezeiza, A, 2013). Some assessment activities that require evaluation tools based on authority are: seminars, workshops, practice exams, interviews, debates, and co-evaluation of activities in cMOOC.

Tools based on social interaction

Undoubtedly, the communication potential of social networks is still largely undiscovered and should be studied more in depth (Guerrero, 2010). Currently, this potential is being introduced in the education system, maximizing the opportunities offered by social networks not only in terms of MOOC, but also as a support tool for traditional classes. Some instruments that fit in this system of social interaction are: anecdotal evidence, portfolios, collaborative Wiki, gamification or motivation based on collaborative games, surveys and questionnaires, chats and forums, projects, workshops, tasks, exercises, activities, and generated knowledge or collaborative learning in cMOOC and xMOOC.

Based on this data, peer review based tools could be placed between social interaction tools and authority-based tools. However, given that a key part of authority-based tools is that the evaluator should be skilled in the subject matter, it would be more accurate to say that peer review systems constitute MOOC assessment tools based on social interaction. Students are peers and therefore cannot be considered authority. Furthermore, in the experience that will be detailed in this study, many of the students commented on the forum about the difficulty involved in evaluating work about something they were learning.

Need for the study

The MOOC high dropout rate makes it necessary to study how to keep those students throughout the course and ensure their learning. But the question that comes into play is whether it is better to decrease the dropout rate or to improve the quality of learning, a question asked by most of the institutions that venture into online teaching. The current abandonment rate of MOOC is hovering around 95%, but this may also be due to the "curiosity" that these online courses are generating. Many students register because they want to know what MOOC is and current statistics do not show this data. It would be interesting to include in the records a checkbox to indicate if the student just wants to try or audit the course. This could improve statistics, at least at these early stages in which MOOCs are giving rise to so many expectations (Acosta, 2013).

Hypothesis

This experiment was performed to determine whether the use of the peer assessment tool is useful or not to reduce the high dropout rate currently experienced in massive online courses.

Methodology

In order to find a valid answer to the question in the hypothesis, a peer review task was included in a MOOC about videogames with more than 7300 students. Thanks to this broad and diverse sample of students (cf., Figures 1, 2, 3 and 4) the possible extension of these results to other massive online courses can be ratified.

The experiment was conducted on a platform where many other Spanish massive courses exist, some with tasks in pairs and others not. Studies on other platforms (Jordan, 2013) show that most MOOC feature self-assessment (usually relegated to a single type of assessment tool, such as multiple choice tests) and/or peer reviews.

- 24 - MCQ (Multiple Choice Questions) and evaluation by peers
- 114 - MCQ (Multiple Choice Questions) only
- 10 – Evaluation by peers only
- 7 – Other

The course was divided into 6 modules with a series of about 10 lessons for each module accompanied by a video for each lesson. The peer review task was introduced in the second module and even though it was mentioned in the initial plan of study or syllabus, several students had not noticed it and therefore they were taken by surprise, which emphasized a decline in the performance of these tasks during the first weeks of the course.

Students were warned that completion and grading would be held during the next two weeks following the beginning of a module and they had to assess at least a student to be graded within the platform. Yet, there were many completion problems because they thought they would have to complete the evaluation until the end of the course.

One of the main comments made by the students in the community of the course concerned how to evaluate peers. Perhaps the assessment rubric was not entirely precise and many options were left to interpretation. It is very important for the rubric to be as specific as possible so that students are able to effectively evaluate their peers.

During week 6 course statistics were recorded to see the evolution of the activity over time and a final survey was conducted. Over 1200 students answered the survey.

Results

Having seen the methodology used in the experiment and that the sample was large enough to refute its reliability, the next step was to detail the most relevant statistical data of the experiment in terms of the peer evaluation tool that was being studied.

First, it seems relevant to compare the completion of the modules among each other, because, as stated above, only one peer review technique was introduced in Module 2. This may give an idea of the difference between this type of assessment and others used in the course, such as multiple-choice questions (cf., Table 1).

It can be observed graphically (cf., Figure 5) how that type of evaluation causes a slight deviation in the completion of the module. Still, it is much more interesting to look closely at this tool within Module 2, because if all modules are mixed, it is possible that other lessons that do not contain peer reviews may mask the statistics of this assessment instrument.

During the last week of the course, substantial differences could be observed between the rest of the lessons and the one which contains the peer review (cf., Table 2), although perhaps more detail can be observed in Figure 6.

From this data, one can already draw interesting conclusions regarding this type of assessment tool in relation to the completion rate of MOOC. Throughout the various stages at which the students had been completing the studied activities, there had been a significant decrease in the completion of peer evaluation activities. Upon completion, this course granted two types of certificates supported by the platform and by the university offering the course. One of them was the certificate of participation, granted to all the students who exceeded 75% of the course, while the other certificate of achievement was given to those who completed 100% of all the activities. Therefore, all those who were unable to complete the peer review task on time were left out of this certificate of achievement, decreasing by far what would have happened if this type of evaluation had not been included. It should be noted that this certificate had cost 40 euros. These fees could fund course costs; therefore a decline could greatly affect the financing of MOOC. With all these results one can get an idea of what students are willing to do in order to complete a massive online course, however at the end of this experiment a survey was sent with different questions about the course and some were highly significant (cf., Figure 7 and 8), these questions were answered by over 1200 students.

Conclusions

From the data obtained in the above study, one can respond negatively to the hypothesis of this experiment. That is to say, the use of peer reviews adversely affects the completion rate of MOOC. This does not mean that learning is of a higher or lower quality, but rather that if the objective is only to increase the completion rate, it is best to avoid these types of assessment instruments.

This experiment also served to improve some course implementation guidelines that are currently being considered for the next version of the course that will begin shortly. For example, peer review activities should be maintained throughout the course as a way to accept and include students who get more interested in the course during subsequent weeks. Many students began at weeks 3 and 4 and therefore had basically no choice to perform the peer review task. Furthermore, platforms should improve this type of assessment instruments. Many of them are not taking into account that some students were not assessed because on some occasions the students who were supposed to conduct the review did not do it. When that happens the task should be given immediately to another student until the work is assessed. It should not happen that students who perform a task are not assessed.

Another point to consider is that the assessment rubric should be very accurate; many students relied on their intuition to assess rather than using the rubric. Many tasks were not properly evaluated because the students were not skilled in the subject matter. The student should take the role of a “robot” that does not know anything and needs to receive all the guidelines necessary to perform a proper assessment. It must be assumed that the student is learning and therefore does not know much. Better than a rubric, the student could receive a small algorithm to be followed step by step to allow careful evaluation of the content, indicating, e.g., what constitutes minimum content, in how many parts content should be divided, what to do if any of the main parts are missing, and what score to assign to each section.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Cuevas, R.F., & Mestaza, M. (2002). La evaluación científica y el sistema de revisión por pares. *CSI Boletín*, *46*.

Ezeiza, A. (2013). ¡Horror! ¡Me evalúa un Robot! *Boletín SCOPEO Nº 85*. Retrieved from http://scopeo.usal.es/horror-me-evalua-un-robot/

Granger, S., Hung, J., & Petch-Tyson, S. (2002). *Computer learner corpora, second language acquisition, and foreign language teaching* (Vol. 6). John Benjamins.

Guerrero, C.S. (2010). Aprendizaje cooperativo e interacción asincrónica textual en contextos educativos virtuales. *Pixel-Bit: Revista de Medios Y Educación*, (36), 53–67.

Human Readers. (2013). Retrieved November 6, 2013, from http://humanreaders.org/petition/index.php

Jordan, K. (2013). Synthesising MOOC completion rates. *MoocMoocher*. Retrieved July 24, 2013, from http://moocmoocher.wordpress.com/2013/02/13/synthesising-mooc-completion-rates/

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale* (Vol. 17). Peter Lang.

Sánchez Acosta, E. (2013). MOOC: Resultados reales. *Elearningeuropa.info*. Retrieved from http://elearningeuropa.info/en/article/MOOC:-Resultados-reales

Steinhart, D. (2000). Summary street: An LSA based intelligent tutoring system for writing and revising summaries. *Unpublished Doctoral Dissertation, University of Colorado*.

Tables

Table 1. Completion Statistics by Module

Design, Organization and Evaluation: Evaluation of videogames and gamification. There were 7,386 registered users		
5689 people started the course and 807 completed it		
Module	**Number of students who started**	**Number of students who finished**
Mo 0. Presentation of the course	4826	4825
Mo 1. History and development of videogames	5373	4211
Mo 2. Designing a videogame	*4004*	*1349*
Mo 3. Roles within the industry	2782	2365
Mo 4. Funding and distribution: The long road	2254	1996
Mo 5. Game review and evaluation. Game as art	1929	1671
Mo 6. Gamification and current trends	1521	1249

Table 2. Module 2 lesson statistics

Activity	**Started**	**Approved**	**Average Grade**
Topic 1: A reasonable doubt	3873	3870	100
Questionnaire: A reasonable doubt	3848	3776	99.683
Topic 2: What is NOT game design?	3833	3831	100
Questionnaire: What is NOT game design?	3773	3728	99.706
Topic 3: Establishing forms	3771	3769	100
Questionnaire: Establishing forms	3705	3662	99.836
Topic 4: What can we do with all this?	3674	3673	100
Questionnaire: What can we do with all this?	3624	3583	99.833
Topic 5: Generating decision making	3602	3599	100
Questionnaire: How to generate decision making	3565	3525	99.887
Topic 6: Let's talk about design with a theoretician: Keith Burgun	3562	3561	100
Topic 7: Levels; the other side of design	3456	3456	100
Questionnaire: Levels; the other side of design	3421	3380	99.941
Topic 8: Miyamoto-San's Master class	3425	3424	100
Questionnaire: Miyamoto-San's Master class	3384	3348	99.91

Activity	Started	Approved	Average Grade
Interview with Raúl Rubio	3378	3376	100
Interview with Lucas González	3209	3208	100
Peer2Peer Activity	*1406*	*1360*	*86.186*
Additional Documentation	2950	2947	100

Figures

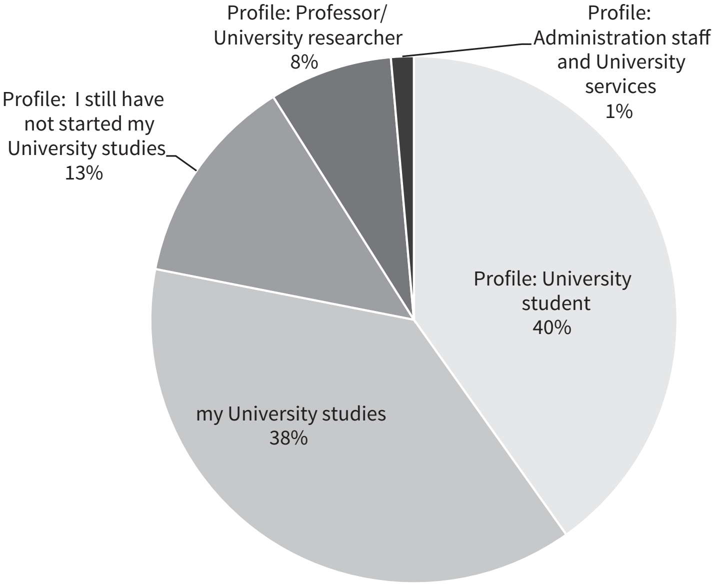


Figure 1. Profile of students to whom the peer review was directed

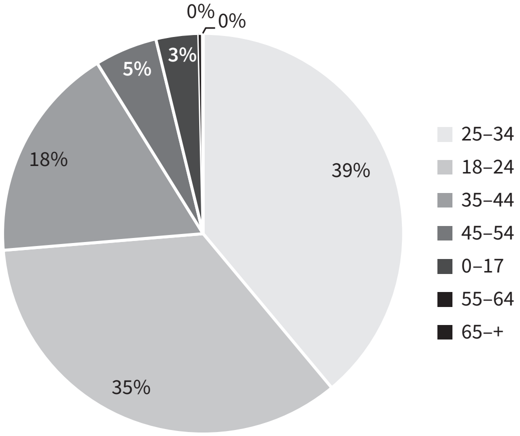


Figure 2. Age of students to whom the peer review was directed

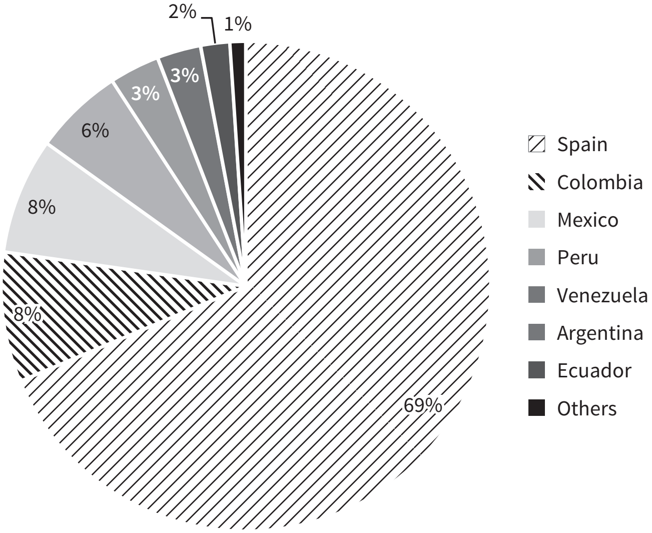


Figure 3. Nationality of students to whom the peer review was directed

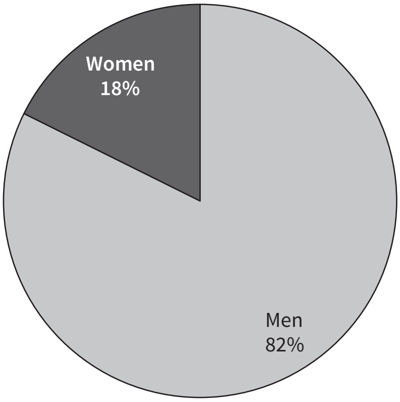


Figure 4. Gender of students to whom the peer review was directed.

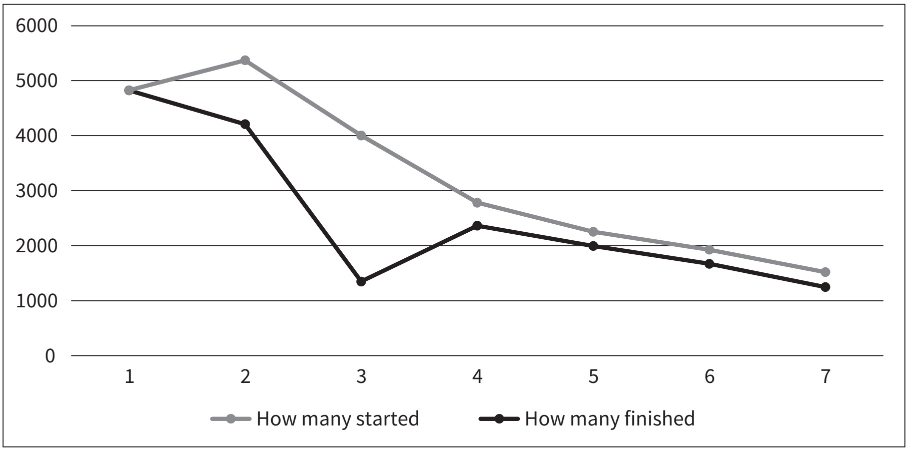
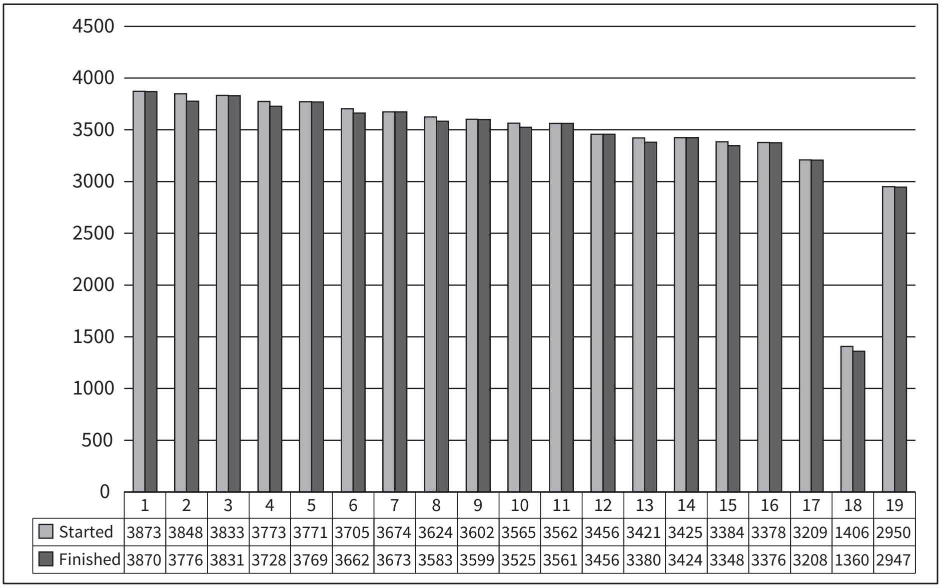


Figure 5. Module completion comparison



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
▢ Started	3873	3848	3833	3773	3771	3705	3674	3624	3602	3565	3562	3456	3421	3425	3384	3378	3209	1406	2950
■ Finished	3870	3776	3831	3728	3769	3662	3673	3583	3599	3525	3561	3456	3380	3424	3348	3376	3208	1360	2947

Figure 6. Peer assessment comparison within the module

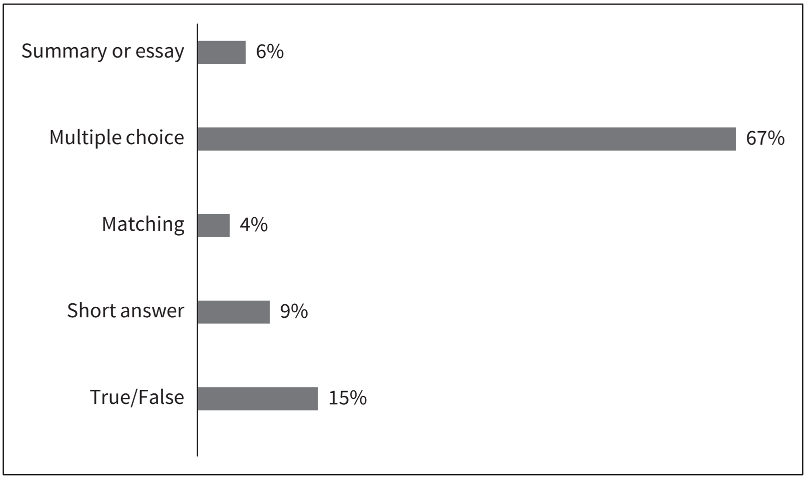


Figure 7. Survey results regarding types of exercises.

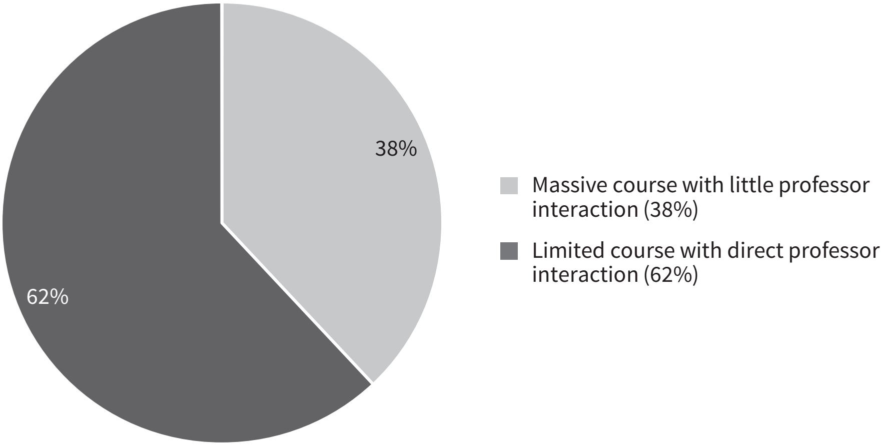


Figure 8. Which types of online course do students prefer?