

Assessment Issues for MOOCs and Large Scale Examinations and Robust, Objective Testing with Reverse Multiple-choice

Indu M. Anand, Sushila Publications, Chelmsford, Massachusetts, USA
Lamia Atma.Djoudi, Synchrone Technologies, Paris, France

Abstract. *For the examinations taken by very large numbers of students, such as the SAT or the ACT, multiple-choice form of questions has traditionally been used. These tests are machine-gradable, uniform, scalable, low-cost, and allow for testing of greater segment of the subject matter and eliminate grader's bias. But the format lacks incisive and probative information for instructors/examiners or reliable feedback for students. Massive Open Online Courses have fueled renewed quest for assessment alternatives, which are discussed. The current approaches leave ample room for improvement, however, especially when students' achievement has to be measured for college credit or certification. The Reverse Multiple-choice Method can provide computerized testing and an elegant answer to assessment concerns for MOOCs, that may be combined with the other approaches and peer grading.*

Keywords: Assessment, MOOC, multiple-choice, on-line education

1. Introduction:

Educational assessment on large scale

Educational assessment on a very large scale has seen renewed interest in the recent years with the advent of Massive Open Online Courses (MOOCs), since by design, MOOCs may have hundreds of thousands of students or more. This unprecedented spread of internet-mediated education has given rise to new problems of scale. Assessment for MOOCs remains an open problem and an active area of research at major universities, e.g., Stanford, Harvard, MIT and Princeton

Reliable assessment and feedback have always been an integral component of education. Whereas the feedback came instantly for an ancient student learning at the feet of a guru, the immediacy between learning and assessment has been eroding with the education of the masses. Assessment on a massive scale is still a key educational activity, but it needs to be managed

remotely, and online for a MOOC.

Standardized tests, mainly in multiple-choice format, have traditionally been used on very large scale, for instance for Scholastic Aptitude Test (SAT) or American College Testing (ACT), in order to compare achievement levels of college bound high school seniors from diverse institutions. But, the testing methods of SAT/ACT cannot simply be extrapolated to MOOCs, and novel assessment approaches have been tried for them in addition including portfolios, blogs, wikis, forum discussions and peer grading. Especially when college credit or certification of achievement hinges on the assessment, these new testing approaches raise concerns for MOOCs and other massive courses, such as governmental or corporate training.

In this paper, we review assessment issues relating to scale, consider the solutions in vogue, and propose how Reverse Multiple-choice Method (RMCM), a scalable, computerized method may be used, to have a cost effective and reliable assessment component however large the number of test-takers.

2. Assessment options for large scale

2.1 Multiple-choice tests for high school

Multiple-choice tests have been used for high school students for one hundred years. There is considerable collective experience and a wealth of test items for multiple-choice tests. SAT and ACT were conducted exclusively via multiple-choice until 2005, and the format is employed extensively, e.g., as part of admission tests for higher education or licensing tests. Examples include Graduate Record Examination, Graduate Management Admission Test, and Law School Admission Test. Multiple-choice format has many advantages: It is objective, machine-gradable, uniform, scalable, and it allows for a greater cross section of the subject matter to be tested and a wider swath of feedback. Multiple-choice tests are also less

expensive to conduct and grade than open-format essays, and provide grades untainted by grader bias.

For significant, milestone examinations, such as the SAT and the ACT, the appropriate passing bar is statistically determined. Substantial research and resources have gone into these exams to interpret the “raw” scores into appropriate “scaled” scores, so that the students can be meaningfully compared to their peers’ percentiles and no group gains an unfair advantage. In theory, such comparison enables colleges to make admission decisions for a student with some assurance of future success in college.

Contrarily, some studies have also shown a lack of strong correlation between the results of these large examination and success in college. Therefore experimentation with test formats continues at the examining bodies and college administrations about the weight to be given to such “standardized” testing.

In 2006, for example, Educational Testing Service (ETS), which administers SAT, added the requirement of an open format essay to be written by the test-takers during the examination. This addition of a new component to the test of English and basic Mathematics raised the total number of points from 1600 to 2400. In March 2015, however, ETS eliminated the essay and returned to old scheme of test and 1600 points. There are many reasons for these changes, several of which are not relevant to our present discussion.

Two points are noteworthy, however: The cost of conducting SATs rose by more than 30% with the introduction of the essay portion; and, the enormous logistical and cost burdens of open format alternatives are at least partly responsible for the popularity of multiple-choice tests for large scale academic exams.

2.2 Assessment options for MOOCs

Assessment remains a tenuous link in the delivery of knowledge and learning via the MOOCs. Unlike the milestone exams like the SAT/ACT which command significant fees, because of the cost burden, using physical testing centers for many thousands of geographically spread out examinees is out of the question for a MOOC. So is centralized grading of open format long or short essays. For MOOC-scale even the use of multiple-choice exams, where a test-taker merely marks correct answers, is fraught with challenges of proctoring and control of plagiarism and cheating.

Assessment for MOOCs, therefore, often relies on novel approaches. These include many forms of

students’ term portfolios, projects, blogs, wikis and forum discussions, and new ways of feedback and grading, particularly in the humanities. The enormous logistical challenge of grading/evaluating these submissions is handled by recruiting the legions of students themselves through “peer grading.”

2.3 Peer grading

Peer grading has emerged as one of the most promising and studied answers to grading dilemma for MOOCs. In broad terms, the mechanism of peer grading works as follows: Each student grades, and has his or her submission likewise graded by a number (about half dozen) of other students under instructor-specified “rubrics” or guidelines, thus freeing up the instructor from the task of actively reading and grading the multitudes of student submissions. The instructor may additionally provide a few graded assignments as samples.

Peer grading process can be a powerful learning activity in its own right, and has been used in some form for decades, e.g. as evaluative component for seminar courses in higher education. It works particularly well for interdisciplinary or exploratory courses, numbers of students notwithstanding. Historically, however, peer grading has largely been used for small, advanced, in-person classes and its extrapolation to massive online courses is not straightforward. In an influential paper, the Stanford team led by C. Piech et al. [5], advocated sophisticated, statistically “tuned” models to enhance peer grading for MOOCs.

Briefly, in a common version of this approach, the scores assigned by the students for a submission are processed statistically to estimate an unknown “true” score with an acceptable degree of confidence. In another version, the statistical processing estimates the grade that the *instructor would have assigned*, given the raw score data and a few instructor graded sample submissions. In these and other variations of tuned peer grading, the computer relieves the instructor from the drudgery of grading large number of papers.

While the advantages of peer evaluation are oft-mentioned, a major driver for its increasing popularity is the benefit of *cost containment*. By farming out assessment tasks to the students substantial cost savings can be realized.

However, the jury is still out on the validity and reliability of peer grading as an assessment tool, and controversy and difficulties remain. Studies indicate that it is possible to craft reliable peer assessment strategies through careful articulation of the rubrics. Cf. Heng Luo et al. [6]. But contradictory studies also exist. In particular, the issue of conflict is a real concern with any grading option, including blogs and forum discussions, which includes a peer grading component: There may be inherent conflict in co-opting for grading the very students being graded.

2.4 Multiple-choice assessment and MOOCs

As noted above, there are many advantages of multiple-choice format: It is objective, machine-gradable, uniform, scalable and low-cost, and it covers a greater cross section of the subject matter for testing, and feedback. Also, significantly, it largely eliminates grader bias.

However, the use of multiple-choice tests for MOOCs presents unique challenges, including the twin imperatives to eliminate plagiarism and cheating while keeping the cost of testing low.

A key additional issue is that traditional multiple-choice format has limited inherent value as a *probative* tool for testing of “summative” or critical skills. The still developing taxonomy of testing for MOOCs recognizes disparate needs of “xMOOC” testing, i.e., the testing of “formative” knowledge based on material previously presented, and “cMOOC” testing, i.e., testing of “summative” knowledge and application skills to new situations. The consensus so far points to the following: Since multiple-choice format generally lacks incisive and probative information for the instructors/ examiners or reliable feedback for students, it can be effective for xMOOC but *not* for cMOOC testing. But, cMOOC testing to assess and give feedback to the learners about critical thinking and application skills is important, especially in case of MOOCs that include a high proportion of non-traditional students.

Thus, at the present time, there remains a need for scalable, *objective* method of grading that can probe learners’ grasp of the subject matter and provide timely feedback. The Reverse Multiple-choice Method (RMCM), an objective, scalable method presented in

the next section is designed to address this limitation of multiple-choice format. We propose the use of RMCM for reliable feedback and/or as a check for validity and reliability of the grades generated by other methodologies.

3. An advantageous alternative to multiple-choice: RMCM

3.1 Structure of RMCM questions

An assessment strategy incorporating the Reverse Multiple-choice Method (RMCM) holds a unique promise for MOOCs, since it offers the objectivity, efficiency and scalability similar to traditional multiple-choice tests, along with a test of knowledge and understanding generally associated with open format. Grading of RMCM questions is possible on computer, thus making it suitable for online courses in most subject areas at several academic levels, *regardless* of the number of students. Furthermore, the method is compatible with other approaches proposed, pursued and discussed in the literature, and it may be used in addition.

RMCM is based on the observation that it is possible to use a multiple-choice question, with its perspective *reversed*, and task an examinee to reveal their reasons for the answer selection in a *brief, succinct* manner thus: Given the answer choices, task the student to modify the “query” so as to make an incorrect answer *correct* for modified question. This basic logic of RMCM is shown in Figure 1.

The typical and distinguishing steps of a typical RMCM approach are as follows:

- prompt examinee to select an answer choice as the correct answer; record the examinee's selection, assign credit for it;
- prompt the examinee to select at least one answer *not* selected as correct, then ask for a follow-up query to which *this incorrect* answer is a *correct* answer;
- match the follow up query against the database of queries for which the selected incorrect answer would be correct
- provide examinee’s score for the question according to examiner’s or administrative policy based on the result: Full credit for perfect matching, zero or negative credit for a total miss and partial credit for a partial match.


3.2 Example 1

Q: The common intelligence quotient (IQ) scale is Normally distributed with mean 100 and standard deviation 15. What proportion of population has IQ scores between 115 and 130?

A. 68%. B. 95% (C) 13.5% (D) 34%

Since in this case the stem has the accepted and unmodifiable information, the system would look for the queries for answer selections and changed interrogatives as shown in Table 2.

Table 1. Structure of a Multiple-choice Question illustrated for Example 1



Query		Putative Answers
Set of presumed facts (Stem of the question, narrative/equation etc.)	Interrogative Sentence (Call of the Question)	Set of Answer Choices (One correct, others incorrect)
The common intelligence quotient (IQ) scale is Normally distributed with mean 100 and standard deviation 15.	What proportion of population has IQ scores between 115 and 130?	(A) 68%. (B) 95% (C) 13.5% (D) 34%

Table 2. Structure of RMCM question for Example 1

Answer Choice	For given query	Changed query (Provided by the Examinee)	For changed query
(A)	Incorrect	What proportion of population has IQ scores between 85 and 115?	Correct
(B)	Incorrect	What proportion of the population has IQ scores between 70 and 130?	Correct
(C)	Correct	--	--
(D)	Incorrect	What proportion of population has IQ scores between 100 and 115?	Correct

3.3 A unique advantage of RMCM: Capturing context for query interpretation

A notable merit of multiple-choice question format is that it captures context of the query *more concisely* than other formats, since it is possible to view each answer choice as adding *contextual* information for the interpretation of the query.

Furthermore, the selection of one answer choice as “correct” over the other choices generally turns on a few syntactic elements, such as, words, phrases, operations, numbers and symbols etc. which contain the key facts. In RMCM terminology, the syntactic elements that make an answer choice correct are

called Fact Objects (FOs). The value of a fact object for which an answer choice is *correct* is called Fact Value (FV) of the FO, a concept akin to assigning a constant value to a variable in algebra. These concepts are elaborated further by examples below.

In Example 1, for instance, there are *two* fact objects, namely the two end points of the interval of IQ scores, since the selection of an answer depends *only* on the values of those end points.

When creating a RMCM question, the examiner specifies fact objects and fact values for *all* the answer choices. The system provides the platform and editorial support for question creation, then uses examiner’s specifications to *automatically* evaluate student answers or flag unexpected student answers

for human evaluation. If a RMCM question is appropriate to the students' academic level and constructed well with plausible confounders rather than irrelevant incorrect answers, the proportion of answers flagged for human evaluation would be relatively small, even for tests taken by massive numbers, such as the MOOCs.

4. Considerations for Using RMCM

4.1 Question creation

Depending on the subject matter being tested, a RMCM question can be framed and the "task" of modifying the query specified in many ways. For each answer choice that the test-taker regards as an incorrect answer to original query, such tasks may include one or more of the following paradigms:

- (i) Identify the fact objects that need to be changed;
- (ii) identify the fact objects to change from a given list;
- (iii) write in the fact objects that need to be changed;
- (iv) write in the fact objects to change from a given list;
- (v) identify the fact values of a fact object that need to be changed;
- (vi) identify the fact values to change from a given list;
- (vii) write in the fact values of a fact object which need to be changed; or
- (viii) write in the fact values to change from a given list.

A task may be specified in simple terms without using fact object/ fact value terminology. For example:

Find the words /phrases/ symbols or other segments of the query which, if they are changed, will make your selected incorrect answer be the correct answer for the changed question.

Or,

Your selected answer is incorrect because at least one query segment has the wrong value; identify which value(s) from the given list should be assigned to the query segment(s) so that your

selected incorrect answer becomes the correct answer for the changed question.

Similar language may be devised for write-in answers for fact objects or fact values.

4.2 Question Answering

For a student who has learnt the subject matter, RMCM questions may be unfamiliar at first but not much harder than answering traditional multiple-choice questions, and possibly easier than answering long or short essay type questions. The RMCM approach strongly encourages the students to:

- (i) focus on closely reading the fact pattern, (ii) critically evaluate the answer choices, and (iii) recognize the critical pieces of information in the given fact pattern on which the answers turn.

The students also must acquire the skill to deconstruct and reassemble a question. But this is precisely the probative or summative information about a learner that an assessment regime seeks for cMOOC testing, and one often desirable for xMOOC testing.

5. Implementation strategies for RMCM

5.1 Machine implementation schemes

When creating the RMCM question, the examiner provides tables such as Tables 2-3 for query matching by the system; the system provides editorial support and intelligent help with question creation in more complex version.

The inputs from students and examiners are received by the system in appropriate fields by any of the known methods in the art. An interesting and useful method is to allow user to highlight a query segment and insert the highlighted character string into the required fact object/fact value/answer fields. For certain subjects or questions, e.g., in the STEM fields, the fact object/fact value tables can be complex, needing much support from the system.

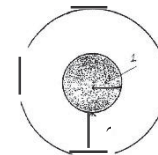
TABLE 3. Answer choice/Fact Object/Fact Value/Scoring for Example 1 of section 2.1

Answer Choice	Fact Object	FO selection score	Fact Value	FV selection score	comments
(A)	[First end point of interval]	50%	85	50%	The two fact objects carry equal weights
	[Second end point of interval]	50%	115	50%	
(B)	[First end point of interval]	50%	70	50%	The two fact objects carry equal weights
	[Second end point of interval]	50%	130	50%	
(D)	[First end point of interval]	50%	100	50%	The two fact objects carry equal weights
	[Second end point of interval]	50%	130	50%	

5.2 Example 2

Julie designed a target computer game. On her computer screen, the circular targets look like the circular areas shown in the accompanying figure, where the radius of the shaded circle is 1 and the vertical length of handle shown in the middle of target is 2. If the computer randomly generates a dot that lands within the circular area, what is the approximate probability that the dot will land in the shaded area?

- A. 1/9 B. 2/9 C. 1/3 D. 2/3

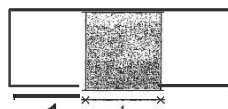


The correct answer is A.

A *Reverse Multiple-choice Question* based on this question would give the following task to the students: Select an *incorrect* answer from the choices A – D. Next, *identify and change* any words, data or other segments of the question, so that your selected answer becomes a correct answer to the changed question.

To automatically score this question, the system would look for the following query changes (possibly on a suitable form):

- A. (No change, but the answer choice should not be selected as an incorrect answer by a student!)
- B. Change vertical length of handle shown in the middle of the target from “2” to length “1.1213”.
- C. Use rectangular figure shown and change every instance of the word “circular” to “rectangular”.
- D. Use rectangular figure shown, change every instance of “circular” to “rectangular” and change “shaded” to “white”.



Example 2 illustrates that the specification of Fact Objects and Fact Values can be made quite robust, even where graphics, or annotated parameters in graphics, may be included within the narrative or stem of an RMCM question. Further a utility to allow highlighting a character string and copying/ inserting it into an appropriate field offers a versatile and flexible way of specifying FOs and FVs for both examiners and students.

In the case here, the RMCM version is not straightforward; there are several ways to change “words, data or other segments” in order to make the incorrect answer become correct for changed question.

In general, the examiner must decide on which answer choices, fact objects and fact values to ask to change based on a given student body, course and target analytical skills to be measured. It may be necessary to provide special answer sheets for graphics.

5. RMCM implementation notes for MOOCs

With the editorial support of the system, the burden of creating RMCM questions may be significantly reduced for an examiner, who can focus instead on creating questions with forethought and deciding what the question will be designed to measure. Since incorrect answers are used to give assessment information, an examiner should use them strategically. An Examiner’s initial expense of time can lead to significant savings of time by automating grading, and substantial cost reduction for MOOCs.

To answer RMCM questions successfully, the students *must* learn to focus on closely reading the facts and paying special attention to incorrect answers in relation to the query. Such learning is useful for both xMOOC and cMOOC evaluations.

6. Using RMCM with peer assessment

We note that for MOOCs, RMCM with its well-articulated test tasks may be provided as part of the “rubrics” for evaluation or as topics of discussion among students in a forum or team. Suitable language of the stem or query, or the nature or purpose of confounding answer choices can serve to focus the students on concrete points of discussion.

A big advantage of using RMCM for peer assessment is that student teams can be recruited to help create suitable question items for Reverse Multiple-choice. The resulting RMCM questions may be submitted to a central “library” or database of queries which the procuring instructor, as well as other instructors/trainers can access. Student teams can help create the questions for

many disciplines, including the STEM fields.

A system of appropriate *rewards* may also be instituted to motivate the students to create test items of lasting value from their experience, and answer questions submitted by other teams/groups. Since a good RMCM question must be answered with critical thought, such a library can be useful for a long time, though an entire test made up of RMCM questions may not be practical to administer in many courses.

In our exploratory classroom experience the introduction of some RMCM questions proved useful as a window into the students’ minds despite automated grading. This provided partial credit where due and a more accurate and meaningful feedback to the student.

7. Conclusion

Reverse Multiple-choice Method incorporates an enhanced and enriched version of multiple-choice format from which it inherits the advantages of machine grading and scalability. The RMCM examiners can draw upon the vast existing resource of multiple-choice questions in all fields.

Our platform for assisting an examiner to create a RMCM *ab initio* or to modify an imported multiple-choice question can provide significant support. As part of further work, we are assembling databases of RMCM questions with suitable query and answer options, which can be useful to teach and evaluate human students, as well as for machine learning. For related research we expect to compare the results obtained by RMCM versus other approaches, and to study the degree to which RMCM can be used to enhance the validity of an approach such as peer grading.

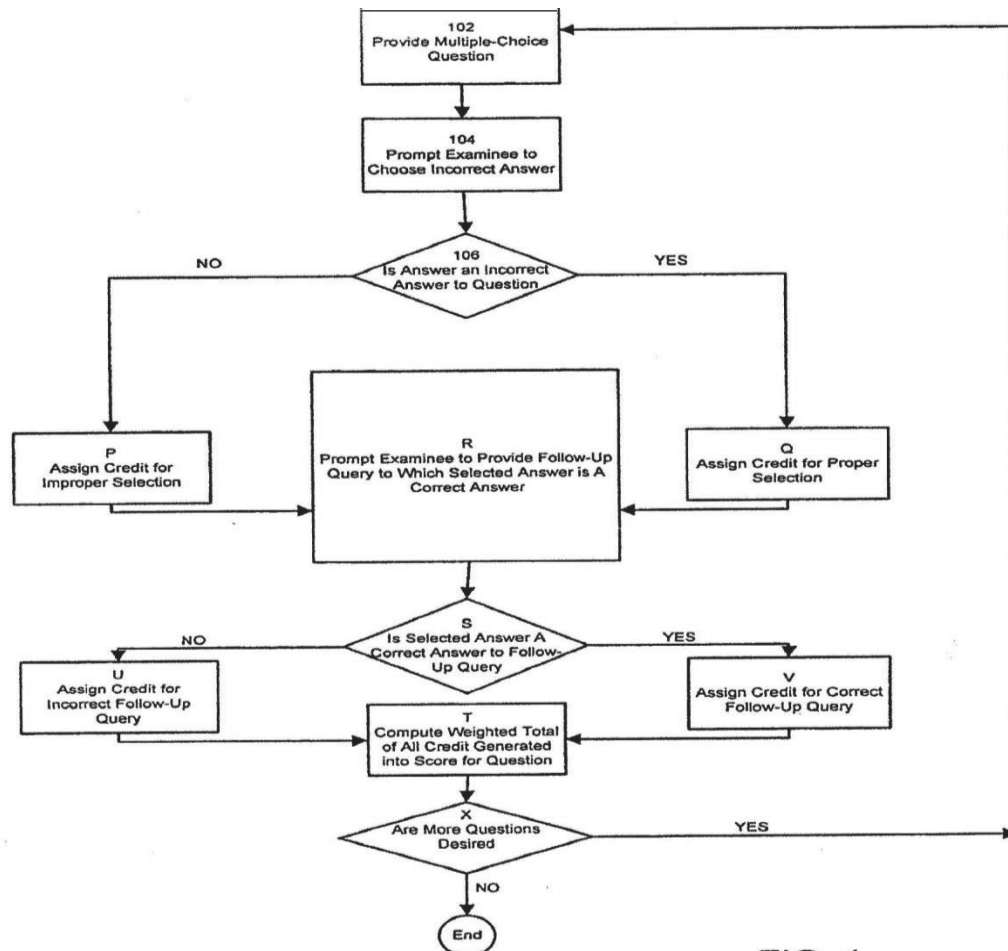


FIG. 1

Fig. 1 depicts RCMC logic. Reproduced from US Patent No. 8,195,085.

8. References

- [1] I. M. Anand, "Method of developing educational materials based on multiple-choice questions," United States Patent No. 7033182, 2006.
- [2] I. M. Anand, "Method of developing educational materials based on multiple-choice questions," United States Patent No. 8195085, 2012.
- [3] P. A. Kirschner, et al., "Do learners really know best? Urban legends in education" *Educational Psychologist*, 48(3), 169-183 (2013).
- [4] Heng Luo, et al., "Is Peer Grading a Valid Assessment Method for Massive Open Online Courses (MOOCs)?" Sloan Consortium 7th Annual International Symposium on Emerging Technologies Online Learning (2014).
- [5] C. Piech, et al., "Tuned Models of Peer Assessment in MOOCs," Stanford University, web.stanford.edu (2013).
- [6] J. A. Reynolds, et al., "Calibrated Peer Review™ assignments in science courses: Are they designed to promote critical thinking and writing skills?" *Journal of College Science Teaching*, 38(2), 60-66 (2008).
- [7] del Sanchez-Vera, et al., "Beyond objective testing and peer assessment: alternative ways of assessment in MOOCs," RUSC, Universities and Knowledge Societies Journal (2015)
- [8] Daniel Thomas Seaton, et al., "Enrollment in MITx MOOCs: Are We Educating Educators?" *Educause Review* (February 2015).
- [9] Hoi K. Suen, "Peer Assessment for Massive Open Online Courses (MOOCs)" *The International Review of Research in Open and Distributed Learning*, 15(3), (2014).