Je-LKS
PEER REVIEWED PAPERS
EMEMITALIA CONFERENCE 2016
Vol. 13, n.3, 2017
ISSN: 1826-6223 | eISSN: 1971-8829
Journal of e-Learning and Knowledge Society
The Italian e-Learning Association Journal

ARE COMPUTER ADAPTIVE TESTS SUITABLE FOR ASSESSMENT IN MOOCS?

Veronica Rossano
Enrica Pesare
Teresa Roselli

Department of Computer Science, University of Bari, Italy
{veronica.rossano, enrica.pesare, teresa.roselli}@uniba.it

Keywords: MOOCs, CAT, Assessment, e-learning

Assessment is one of the basic issues in both formal and informal educational contexts. Current online courses are massive and online, thus it is important to find new strategies to improve the effectiveness of evaluation. In MOOCs (Massive Open Online Courses), indeed, there is also the certification of knowledge and skills acquisition that requires more formal and trustworthy methods. Researchers should work to combine pedagogical and technological solutions to guarantee the effectiveness of learning measures. In this context, the Computer Adaptive Testing (CAT) could be useful to better measure the knowledge acquisition using a quiz, as usually happens in MOOCs. These are the premises of this research work that, to understand if CAT could be suitable for assessment in MOOCs, proposes a first algorithm to measure the acquired knowledge using a quiz-game. The pilot study attests the users' appreciation.

for citations:
Rossano V., Pesare E., Roselli T. (2017), *Are Computer Adaptive Tests suitable for assessment in MOOCs?,* Journal of e-Learning and Knowledge Society, v.13, n.3, 71-81. ISSN: 1826-6223, e-ISSN:1971-8829
DOI: 10.20368/1971-8829/1393

1 Introduction

In education, the assessment of learning is important as well as the instructional process, since it should help both students and teachers in improving the learning-teaching process and, thus, to improve learning effectiveness. The spread of Information and Communication Technologies pushes through different and more innovative assessment procedures that could be helpful in reaching this objective. In this view, it is necessary that also the assessment procedures evolve themselves to respond better to the new needs and features of the users but also of the technologies available, in terms of hardware and software.

Moreover, the spread of new models for online education require as well new learning assessment models. In particular, the assessment process in MOOCs (Massive Open Online Courses) (Kennedy 2014; Liyanagunawardena *et al.*, 2013), that are didactic contents for higher education characterized by unlimited participation and open access, need some revisions. The size of the audience, in fact, requires more specific processes or tools to assure the effectiveness and high quality of learning assessment.

The digital assessment in MOOCs is an emerging challenge. As stated in (Bayne & Ross, 2013) the assessment in MOOCs is an area of great interest and active experimentation. The issues are several, the most important in our view are related to the user's authentication during the assessment and the way useful to effective learning assessment. The first issue has been efficiently resolved in the Eduopen platform[1], that is a network of several Italian Universities that offers free MOOCs. The students can acquire the certification in three different ways. The first one is an attendance certificate, that does not require any payment, the second is a verified certificate, that requires that the final exam should be done in a NICE (Italian Exams Center Network) center where there are examiners that can certify the student's identity; the last is the traditional (i.e. in presence) examination at a University of the network and allows the attribution of ECTS (*European Credit Transfer System*).

As regards the assessment issue, there is a lively discussion. Different solutions have been studied from both pedagogical (self-assessment, peer assessment, co-assessment) and technological points of views.

In this context, the objective of the research herein presented try to address this issue investigating about a new approach to digital assessment. The Computer Adaptive Tests (Sands, 1997) are a methodological and technological solution that could improve the effectiveness of evaluation processes when they involve a wide range of subjects that cannot be classified a priori. To become familiar with the CAT, the researchers have defined a first approach of this new

[1] https://learn.eduopen.org

methodology in a quiz-game. The paper presents the algorithm defined and a first pilot test of the implemented quiz-game.

2 Learning Assessment: technological solutions overview

The learning assessment in online courses is always one of the hot topics in educational research field. In the context of MOOCs, the learning assessment is a thorny issue, since a certification could be acquired by the student (Clements & Cord, 2013). So, it is important that the procedure adopted in the evaluation should be trustworthy to certificate the knowledge and skills acquisition.

In the literature, solutions have been proposed in several directions. (Admiraal *et al.*, 2015; Muñoz-Merino *et al.*, 2015), for example, consider the integration of the qualitative measures that allow the teacher to have more complete feedbacks to certify the knowledge and skills acquisition. The proposed measures include the interaction with didactic resources and activities. Other researches use the Learning Analytics and Learning Dashboard (Pesare *et al.*, 2015; Siemens, 2012; Knight *et al.*, 2014) that allow users to better visualize the monitoring data of virtual learning environments and to infer informative results about the learning process. Some other solutions are focused on designing and implementing tools able to simulate the presence of the teacher or tutor to provide personalized feedbacks to students (Huertas, 2011).

Moreover, the trend is to design e-assessment systems that allow to have different tools and data useful to improve the quality of certification of skill and knowledge acquisition (Crisp, 2007). Sitthiworachart et al. (2008), in this view, propose a novel framework for e-assessment which captures the essential success factors. Another useful strategy could be the Computerized Adaptive Testing (CAT) (van der Linden & Glas, 2000; Wainer *et al.*, 2000). The adaptive tests can adapt the difficulty level of the questions to the student's ability level.

In this view, the authors research aims at evaluating this technological approach in order to adopt it as one possible solution in the context of MOOCs. For this reason, a first study has been conducted to better understand the technique and a quiz-game has been developed.

3 Computer Adaptive Test

The Computer Adaptive Test is a computer-based test that adapts the type and sequence of the questions to the examinee's ability level. In other words, a CAT is a dynamic test that is built by selecting the question to be posed from a set of possible questions (item bank) according to the answers provided by the student during the evaluation process as well as a human teacher does. In

this view, the CAT allow more effective quantitative tests to be implemented than those created with the Classical Test Theory (CTT). Standard fixed tests pose the same number of questions to all users and measure the knowledge acquisition by means of the transformation of the number of correct responses in a score. In this case, the assessment is related to the sample of questions and, for these reasons, they could be less informative in context in which the population is wide and not classifiable a priori, as the MOOCs users are. Moreover, in a traditional test not all questions are necessary to assess student's learning; the questions may be too easy or, on the contrary, too hard and so the test result may provide poor information about the student's actual learning gain. The adaptive tests, instead, build the questionnaire during the test session picking only those questions that are necessary to assess the specific student. Usually, the first question has a medium level of difficulty, the level of difficulty of the next is defined on the basis of the student's answer. The examinee's ability level and the level of difficulty of the questions are defined dynamically during the interaction and it stops when the obtained value is recognized as the best result. This process allows to build tests with different length. Indeed, one of the main objectives of the CAT is to measure the knowledge acquisition level using the minimum number of questions. Thanks to the use of CATs it is possible both to reduce the time required to answer the test, by dynamically selecting the question to be posed, and to maximize the precision of the learner' s evaluation. Thus, a CAT poses easier questions to low ability learners and harder questions to high ability learners. The scoring is calculated on the basis of the learner's ability and the item difficulty. For example, if two students answer the same number of questions, the one who answered the hardest questions will have the highest score. For these reasons a CAT could be less boring for students, because high ability learners do not have to answer useless questions (too easy for her/his ability) and, on the contrary, low ability learners do not have to try to answer too difficult ones.

3.1 Item Response Theory (IRT)

As already mentioned, the main objective of a CAT is to define dynamically the composition of a test by adapting the type and the sequence of questions to the user's capabilities. In this process, it is clear that to build a CAT a prior classification of the items and the learner abilities is needed. The psychometric literature proposes different approaches, the most widely used of which is the Item Response Theory (IRT) (Baker, 2001; Hambleton & Swaminathan, 1985). The Item Response Theory, also known as latent trait theory, is a statistical approach used to define the probability that a user can answer correctly to a specific item. The probability is calculated using the user's ability level and

some item parameters. In other words, the IRT is based on the relationship between individuals' performances and the ability that item was designed to measure (Baker, 2001; Hambleton & Swaminathan, 1985). The user's ability level, or *latent trait* generally denoted using θ, represents the latent variable, a variable not directly observed but inferred through a mathematical model from other variables. The latent trait has impact on the subject's performance. Moreover, the subject's performance is influenced also by some psychometric features of each item: the item *discrimination* parameter (a), the item *difficulty* parameter (b), and the *guessing* (c).

In the literature, there are different IRT models that could be classified on the basis of the number of abilities measured by the test (unidimensional or multidimensional), and on the basis of the number of considered parameters (1, 2 or 3 PL). The 3 PL uses all three parameters, the 2PL assumes that no guessing influences data, and the 1PL assumes that guessing is embedded in the ability and all the items have the same discrimination, thus the P(θ) is calculated using only the difficulty as parameter. For the purpose of this study that is aimed at evaluating the applicability of the adaptive testing theory to the MOOCs context, we choose to apply the 3PL model to calibrate the item bank.

3.2 The Three-Parameter Logistic Model

The calibration is the basic process needed to define an item bank in which each question is useful to measure the subject's ability. In the calibration process one of the above-mentioned models (1, 2, or 3 PL) could be used. In this context, a Three-Parameter Logistic Model (3PLM) was preferred in which the probability that the item *i* could measure the latent trait θ is defined as in (1).

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \qquad i = 1,2, \ldots k \quad (1)$$

For each item *i* the Item Characteristic Curve (ICC) can be defined. It represents the relationship between the probability of correct response to an item and the ability scale. The P(θ) will be small for examinee of low ability and large for examinee of high ability (Baker, 1985). Then, in order to allow the algorithm to select the right item to submit, the Item Information Function (IFF) represents the range of subject ability that the item *i* is able to measure. The analysis of IIF is important to choose the most informative item for a specific range of subject ability. For example, an item is informative for a subject with ability level between 1.0 and 3.0 if the maximum of IFF is reached in θ = 2. To define a high-quality test, each item should be submitted to several processes

of elaboration, revision, experimentation and validation. Because of this, in our research we applied the CAT approach to a quiz-game to evaluate the feasibility.

4 A quiz-game using CAT

To figure out if the CATs could be useful in MOOC assessment, we have applied the adaptive testing approach in two quiz games. The first one (non-adaptive) was aimed at calibrating the bank items, the second was aimed at define an algorithmic strategy to create an adaptive test. Moreover, the second test was used to evaluate the users perceived usefulness. The questions are related to general knowledge, such as history, geography, literature, and science. The games are addressed to subjects of 15 years or more.

4.1 The non-adaptive quiz-game

To define a CAT a first edition of a non-adaptive quiz-game was defined. It aimed at gathering and preprocessing data about the item bank. In other words, this first edition was launched online to collect answers from about 200 subjects to 30 questions. The game simulates a competition between the computer and the user. The idea to implement the game-quiz as a competition was only to make the quiz fun for the users. The final goal of this game was to submit the questions and collect the answers. The competition against the system was only used in order to push the users to play at least twice. At each step the user can choose the subject and the difficulty of the question to be posed to the antagonist. At the end the scoring is calculated as described in the following sections and the winner is celebrated.

4.2 Data processing

The non-adaptive quiz game allows answers of the 30 questions to be gathered. For each item, about 150 answers were collected. The data was used in the calibration process using the 3PL described in section 3.3. The first step is to define the parameters (discrimination, difficulty and guessing) of each item. There are several software tools that allow to define these parameters starting from the subjects' answers. We used Param3PL[2], that is a public domain, freeware tool for calibrating items and individuals using the 1 and 3 parameter logistic item response theory models. As an example, table 1 shows the results of the calibration process of the first 4 items.

[2] http://echo.edres.org:8080/irt/param/

Table 1
OUTCOME OF PARAM3PL SOFTWARE

Item	Parameter a (Discrimination)	Parameter b (Difficulty)	Parameter c (Guessing)
1	0.376	-1.829	0.000
2	0.166	0.146	0.000
3	0.786	-0.711	0.104
4	0.134	5.000	0.416
...	...	...	...

Those values have been used to calculate the ICC and the IIF of each item. Moreover, for each item in table 2 are reported the max value of the IIF and the ability level in which the maximum is reached. This is useful in the selection process since, for example, item 4 will be selected only if the subject's ability is approximately 3. It will be not informative for those subjects whom ability level is between 1 and -5.

Table 2
ITEM INFORMATION FUNCTION (IFF) FOR EACH ITEM

Item	MAX (IIF)	Level of ability in which the MAX is reached (IIF)
1	0.035	-2
2	0.006	0
3	0.120	-1
4	0.259	3
...	---	....

The ultimate outcome of the calibration is the calculation of the $P(\theta_m)$, i.e. the probability that a subject with *m* ability level can answer correctly to the specific item. In table 3, there is an example of the different value of $P(\theta)$ for a question that is classified as difficulty level 0. This means that a subject with ability level equal to -4 has a probability of 0.36 to answer correctly, whereas for a subject with ability level equal to 2 the probability is higher (0.61).

Table 3
PROBABILITY OF CORRECT ANSWER FOR EACH ABILITY LEVEL

Question Difficulty	$P(\theta_{-5})$	$P(\theta_{-4})$	$P(\theta_{-3})$	$P(\theta_{-2})$	$P(\theta_{-1})$	$P(\theta_0)$	$P(\theta_1)$	$P(\theta_2)$	$P(\theta_3)$	$P(\theta_4)$	$P(\theta_5)$
level(0)	0.32	0.36	0.4	0.44	0.48	0.57	0.6	0.61	0.65	0.66	0.7

4.3 The adaptive quiz-game

The quiz-game, named QuizMania, was designed and implemented in order to be used by subjects of 15 years or more. The interaction, graphic and navigation was defined to be the easiest as possible. Moreover, since the test was implemented as a quiz-game, a score and a leaderboard were defined. The score is calculated according to the adaptive testing approach. The score for each item, indeed, is increased or decreased on the basis of the user's ability level. Thus, assuming that a user with ability level equal to *m* gives a correct answer for the question that has the level of difficulty equal to *n* the score is calculated as following:

- $1 + (1 - P(\theta_m))$, if the answer is correct
- $-1 * P(\theta_m)$, if the answer is incorrect

where $P(\theta_m)$ is defined as in table 3. This strategy has been defined since it reasonable that if a low ability user gives the right answer to a question that is higher for the ability level defined by the game, a reward should be given. Moreover, at the increasing of the user's ability the probability that s/he gives the right response is higher, then the rewarding will decrease, and the incidence of any wrong answer will increase. At the end of the quiz-game, a leaderboard is displayed in which the players with the best scores are listed.

5 Users Test

To evaluate the users perceived usefulness and usability a pilot test was conducted. The sample was composed of 40 users. It was impossible for us to make a classification of the users since it was distributed using the author's Facebook page. Moreover, since the game was designed for a wide range of age and the questions were concerning general cultural aspects a classification of the users could be non-informative for the goal of the research. The questionnaire submitted was composed by 20 multiple choice questions using a 5-likert scale for the answers. Some of them aimed at measuring the usability of the quiz-game, some aimed at measuring the user's perceived reliability of the adaptation process.

For what concerning the first group of questions, positive results have been obtained. The 80% of users express high appreciation about both the navigation directions and the 85% of them stated that the language used was easy to understand. There were some doubt concerning the colors used (too dark for someone) and the graphics elements, someone said that too cartoon style was used (Figure 1). This is due to the wide range of age of intended users. A profound restyling of the graphic aspect is needed. The same reasons could explain the results in Figure 2 about the fun and entertainment dimension.

Moreover, a quiz-game is not the funniest form of entertainment.

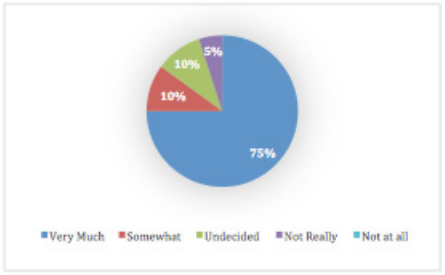


Fig. 1 - Are colors and graphic pleasant?

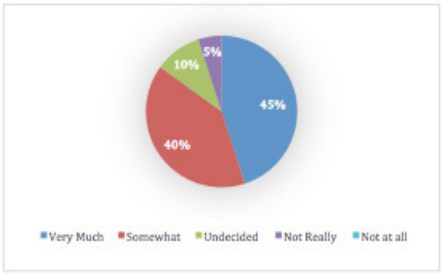
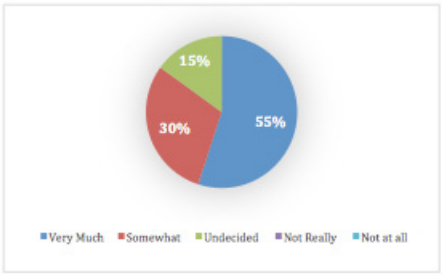


Fig. 2 - The game is fun?

As regards the user's perceived reliability of the adaptation process, 34 users agreed with the proposed classification of questions, only 6 did not give any judgement (Figure 2). For what concerning the score obtained, the majority of the sample stated that the obtained knowledge assessment was acceptable, thus they think that the system is able to measure the knowledge owned by the users (Figure 3).



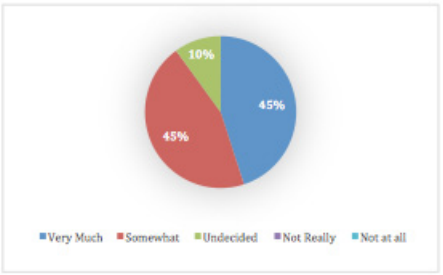


Fig. 3 - (1) The classification of questions represents the actual difficulty of the questions; (2) The system is able to measure the knowledge

Conclusion

The assessment is one of the trend topics in education, both from pedagogical and technological point of views because of the spread of online courses. Moreover, the adoption of new models for e-learning, as MOOCs are, requires more reliable methods to assess the knowledge acquired using the didactic contents. Because the massive peculiarity of MOOCs, multiple choice quizzes are the most largely tools used in final exams. Thus, it is necessary to use innovative algorithmic strategies that allows to differentiate the learning

measures. The paper proposed to use a CAT that allow to dynamically compose a test according to the user's ability. This prevents also user's frustration in answering quiz that are too difficult for their specific profile. A first prototype of a CAT was implemented using a quiz-game, and a first experience about the definition of an algorithm to assign a score was used. The appreciation of the users allows us to go ahead in this research. The next step will be to apply this approach to a MOOC.

Acknowledgments

We would like to thank Maria Magliulo and Elisabetta Cascione who designed and implemented the game quizzes in their thesis.

REFERENCES

Admiraal, W., Huisman, B., & Pilli, O. (2015), *Assessment in Massive Open Online Courses*. Electronic Journal of e-Learning, 13(4), 207-216.

Bayne, S., & Ross, J. (2013), *The pedagogy of the Massive Open Online Course: the UK view*. York, UK: The Higher Education Academy.

Baker, F. (1985), *The Basics of Item Response Theory. CHAPTER 1 The Item Characteristic Curve*. Retrieved from http://echo.edres.org:8080/irt/baker/

Baker F.B. (2001), *The basics of item response theory*, ERIC Clearinghouse on Assessment and Evaluation, Washington, available at http://ericae. net/irt/baker.

Clements, M. D., & Cord, B. A. (2013), *Assessment guiding learning: developing graduate qualities in an experiential learning programme*. Assessment & Evaluation in Higher Education, 38(1), 114-124.

Crisp, G. (2007). *The e-assessment handbook*. Continuum.

Hambleton, R. K., & Swaminathan, H. (1985), *Item response theory: Principles and applications (Vol. 7)*. Springer Science & Business Media.

Huertas, A. (2011), *Ten years of computer-based tutors for teaching logic 2000-2010: Lessons learned*. In Tools for Teaching Logic (pp. 131-140). Springer, Berlin, Heidelberg.

Kennedy, J. (2014), *Characteristics of Massive Open Online Courses (MOOCs): A Research Review*, 2009-2012. Journal of Interactive Online Learning,13(1).

Knight, S., Shum, S. B., & Littleton, K. (2014), *Epistemology, assessment, pedagogy: where learning meets analytics in the middle space*. Journal of Learning Analytics, 1(2), 23-47.

Liyanagunawardena, T. R., Adams, A. A., & Williams, S. A. (2013), *MOOCs: A systematic study of the published literature 2008-2012*. The International Review of Research in Open and Distributed Learning, 14(3), 202-227.

Muñoz-Merino, P. J., Ruipérez-Valiente, J. A., Alario-Hoyos, C., Pérez-Sanagustín, M., & Kloos, C. D. (2015), *Precise Effectiveness Strategy for analyzing the effectiveness*

of students with educational resources and activities in MOOCs. Computers in Human Behavior, 47, 108-118.

Pesare, E., Roselli, T., Rossano, V., & Di Bitonto, P. (2015), *Digitally enhanced assessment in virtual learning environments*. Journal of Visual Languages & Computing, 31, 252-259.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997), *Computerized adaptive testing: From inquiry to operation*. American Psychological Association.

Siemens, G. (2012), *Learning Analytics : Envisioning a Research Discipline and a Domain of Practice*. 2nd International Conference on Learning Analytics & Knowledge, (May), pp.4–8. Available at: http://dl.acm.org/citation.cfm?id=2330605.

Sitthiworachart, J., Joy, M., & Sutinen, E. (2008), *Success factors for e-assessment in computer science education*. In E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (pp. 2287-2293). Association for the Advancement of Computing in Education (AACE).

van der Linden, W. J., & Glas, C. A. (Eds.). (2000), *Computerized adaptive testing: Theory and practice*. Dordrecht: Kluwer Academic.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000), *Computerized adaptive testing: A primer*. Routledge.