A Method for Generation of Multiple-Choice Questions and Their Quality Assessment

Aleksandr Saiapin

Siberian State University of Science and Technologies, Krasnoyarsk, Russia
alstutor@gmail.com

Abstract. The purpose of this study is to propose a reliable multiple-choice or multiple-response assessment generation procedure. A question of the assessment passing threshold selection is considered. The study uses simulation for the test outcomes distribution evaluation which is used for the assessment passing threshold calculation. The relation between the number of answers and distractors in a task and assessment reliability is shown. The reliable assessment generation procedure based on simulation and statistics is proposed. The web application for students testing purposes based on the proposed approach is implemented. The proposed method can be used for online studying as well as offline ones.

Keywords: Multiple-choice assessment · Multiple-response assessment · Online assessment · Simulation modeling

1 Context

Online multiple-choice tests, as well as multiple response tests and questions, show significant growth last few years due to the wide spreading of massive open online courses [2, 9]. It becomes even more important at the moment because of the coronavirus pandemic [6].

In the case of a remote learning situation, it is important to develop a cheating-proof and fair assessment procedure.

The vital parameter of multiple-choice tests and multiple response questions is their discrimination abilities, in other words, the ability to divide the testees into two groups, the ones who have enough knowledge, and the ones who don’t. It is also good to grade testees, to evaluate the knowledge level of a testee in comparison to the others. There are a lot of works devoted to designing good in some sense assessments [8, 11].

Using the percentage of correct answers given by a testee during the assessment procedure as a measure of testee’s knowledge, abilities and skills require a method to set up the correct threshold that divides testees into the ones who successfully passed the assessment and the ones who don’t.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2021
M. E. Auer and T. Rüttmann (Eds.): ICL 2020, AISC 1329, pp. 534–543, 2021.
https://doi.org/10.1007/978-3-030-68201-9_52
2 Purpose

The main aim of an assessment of a testee is to determine the abilities of the testee we assess to act at a given level of Bloom’s taxonomy [5]. As it was already mentioned, the assessment is a means to divide the testees into the two groups: the ones who have the required abilities, and the ones who haven’t. It is also good to have the means to compare the abilities of the testees who passed the assessment, in other words, to range the testees according to their knowledge, skills, and abilities level.

It is also important to prevent cheating during an assessment. According to if we perform the assessment online or offline using computer means, the ways to prevent cheating can be different, but there are means to reduce the cheating no matter which kind of assessment we use.

It’s obvious that passing any multiple-choice question, as well as multiple response questions, is a probabilistic procedure, and both false positive and false negative results are possible. However, it doesn’t mean that all the events (true positive assessment result, true negative assessment result, false-positive assessment result, false-negative assessment result) have to have equal probabilities, and the main feature of a good assessment is that the first two events have to be much more probable than the other two.

The goal of the work is to describe a way for an automatic generation method for multiple-choice assessments and multiple response assessment as well as their quality assurance allowing to generate assessments with a given discrimination ability, and an ability to range the testees who successfully passed the assessment according to their knowledge, skills and abilities level.

3 Basic Terms

The main aim of an assessment, either it is multiple choice or multiple response test, is to evaluate testees’ knowledge, skills, and abilities according to Bloom’s taxonomy. It includes six levels [5]:

1. remembering: retrieving, recognizing, and recalling relevant knowledge from long-term memory.
2. understanding: constructing meaning from oral, written, and graphic messages through interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining.
3. applying: carrying out or using a procedure for executing, or implementing.
4. analyzing: breaking material into constituent parts, determining how the parts relate to one another and to an overall structure or purpose through differentiating, organizing, and attributing.
5. evaluating: making judgments based on criteria and standards through checking and critiquing.
6. creating: putting elements together to form a coherent or functional whole; reorganizing elements into a new pattern or structure through generating, planning, or producing.
It is obvious that an assessment hardly can be used at the highest levels of the taxonomy at least because it’s hard to deal at these levels without formulating consistent oral or written answers, not mentioning creation or proposing some practical solutions.

A multiple-choice test is a form of an objective assessment in which respondents are asked to select the only correct answer out of the choices from a list [7] while a standard type of multiple-response or multiple-answer question looks like an MCQ except that the student can choose more than one answer.

A typical task includes a stem (the text that describes the task itself, may include diagrams, pictures, photos), options (the options for testee to choose from, may also include diagrams, pictures, photos) which consist of answers (the options to be marked as correct) and distractors (the options not to be marked). The options chosen by the testee are called the response.

The components of the task of an assessment are shown in Fig. 1.

![Fig. 1. Task structure](image)

An assessment consists of a list of tasks. The number of the tasks in an assessment may vary, but typically an assessment consists of about 10–15 tasks.

4 Test Quality

As the goal of a test is to estimate the knowledge, abilities, and skills of a testee, a test must provide special features.

These features are:

1. validity
2. discrimination ability
3. reliability

Each feature supports assessment results in different ways.

The validity of an assessment means it estimates exactly what it is intended to estimate, or, in other words, validity is defined as the extent to which scores obtained on an assessment instrument represent true knowledge [12]. The validity of a test is defined by its content (the content of the stems and options of the tasks of the test) mostly and is set at the moment of the test tasks’ development.
The discrimination index of a test shows how it differentiates the testees who have required knowledge, skills, and abilities from the ones who haven’t.

The Classical Test Theory describes the reliability of a test in terms of observed score, true score, and random error component. It states that an examinee’s observed score \((X)\) can be decomposed into her/his true score \((T)\) and a random error component \((E)\) \((X = T + E)\) \((3)\). In fact, the less \(E\) value is, the more reliable assessment we use.

In some way, we can perceive the reliability of a test how its ability to maintain the same (or close enough) results for the same testee from an attempt to attempt. For example, if the testee passes the test a few times in a row, the results of the test would be the same (at least in terms of passing/not passing the test). In the reality the question of the reliability of a test is not that simple: the testee who passes a test few times in a row, learns from his or her own mistakes, so theoretically it is possible that the testee who had not enough knowledge before the test, would get it during the first attempts of passing the test.

The question of the reliability of a test is connected with the process of test development as we are going to see later in this article.

5 Test Result Calculation

There are a lot of ways to calculate the results of a multiple-choice assessment or multiple response assessment. In this article, we consider the approach the author uses for his own testing purposes.

The result of a test is calculated according to the following formula:

\[
S = \frac{|A \cap M| + |D \cap N|}{|A \cup D|}
\]

where \(S\) is the score of the assessment,

- \(A\) - a set of a tests’ answers,
- \(C\) - a set of the tests’ options, marked by a testee,
- \(D\) - a set of a tests’ distractors,
- \(N\) - a set of tests’ options, not marked by a testee.

In some way, this metric corresponds to the Jaccard index \([10]\), also known as Intersection over Union or the Jaccard similarity coefficient.

The formula here is modified a bit according to the fact that the answers set and the distractors set have no common elements.

Using the formula allows us to avoid the situation of cheating by marking all the possible options of a test (maximizing the number of positive answers), or not marking options at all (minimizing the number of mistakes).

This formula is applicable both for multiple-choice assessments and for multiple response assessments.
6 Test Reliability

6.1 Passing Threshold Calculation According to the Stochastic Nature of Test Passing

The main question of an assessment is how to decide whether the test is passed or not, according to the testee’s score, calculated using the formula provided earlier.

In most cases the author of a test sets the passing threshold, different authors set passing thresholds different ways. Typically the required percent to pass a test is set to 60–75%.

It is obvious that the passing of a test is a stochastic procedure. A testee may pass a test having no knowledge, skills, and abilities at all, though the probability of that event is reasonably low.

We need to establish a procedure to calculate a test passing threshold that guarantees that a test cannot be passed without having knowledge, skills, and abilities with the given probability. In other words, the threshold value has to be chosen the way that in 95% (for example) attempts a testee without knowledge, skill, and abilities fail the test.

The intention to make an assessment as cheating-proof as possible often gets tests authors to provide different tests for the different testees (generated from a common task pool), which makes it even harder to set the passing threshold.

To establish the correct passing threshold for a given test, we need to know the probability of each test outcome. To get the probabilities we can perform test passing simulation. It gives us the following pictures.

Here the distributions for different answers and distractors numbers in an assessment are presented.

The X-axis shows a test outcome calculated using the formula we discussed earlier, and the Y-axis shows the probability of the given test outcome. A is the number of answers in a task, and D is the number of distractors. All the tasks in an assessment have the same number of answers and distractors. The first column of the diagram corresponds to multiple-choice assessments, the second one corresponds to multiple-response assessments. The simulations for each answer/distractors ratios were performed for 100000 times.

As you can see, the greater numbers of answers are more preferable as they produce more possible test outcomes. Also, the probability to occasionally get higher outcomes is lower than for a lesser number of answers in tasks.

To set the passing threshold we can use the following method:

1. set the probability of occasional test passing (5% for instance)
2. starting from the highest outcomes sum the probabilities of the outcomes
3. stop when the sum is equal or higher than the required test passing probability we set
4. the outcome we found is the threshold we are looking for.

The vertical lines on the diagram above show the corresponding tests passing thresholds (Fig. 2).
In real-life scenarios, the number of answers and distractors are seldom the same for all the tasks in a test (unless the authors of a test specifically intended to create the assessment that way). In most cases, the number of answers and distractors differs from task to task so we have to check that kind of scenario too.

It is obvious, that for a reliable test the higher outcome is, the harder it is to get it answering the test tasks by random.

Unfortunately, it is not correct for the tests where the number of answers and distractors differs for different tasks of the test.

Distributions for real tests are shown below (Figs. 3 and 4).
The first column demonstrates the whole distribution picture, the second one shows the part of the distribution that corresponds to the passing test outcomes, and the last one pictures the detailed part of the distribution where the rule described above is broken.

It is obvious, that the shown assessments are not reliable in the meaning we discussed earlier, as they provide a higher score with a higher possibility to achieve it comparing to the probability of the lower outcomes.

So we as an assessment developers have to establish the procedure to generate an assessment that is reliable, the way so the probability for a testee to get a high outcome is as lower as higher the outcome is.

Also, if we are to grade the testees, the number of assessment outcomes that correspond to a successful assessment passing has to be as high as possible.

So finally we propose a reliable assessment generation procedure.
6.2 A Reliable Test Generation Procedure

To achieve all we have mentioned above the following approach is proposed:

1. Set the probability of occasional test passing (5% for instance) \( P \)
2. A given number of tasks for the assessment are extracted from a task pool (what allows us to reduce the cheating probability during the assessment, generating a unique assessment for each testee);
3. Then for the assessment we generate, the procedure of random answering is preformed for the fixed number of times, simulating the assessment passing process. It allows us to discover the distribution of the outcomes of the test.
4. The threshold is defined as a percentage of correct answers, that provides the probability to pass the assessment that is less or equal to the given value of probability \( P \):
   1. starting from the highest outcomes sum the probabilities of the outcomes.
   2. stop when the sum is equal or higher than the required test passing probability \( P \).
   3. the outcome value we found is the threshold \( T \) we are looking for.
5. For each outcome that is higher than the threshold \( T \) check if the probability to get each next assessment outcome is lower than the current one is. If it is not so, the assessment has to be rejected.
6. Then the ranging ability of the assessment is evaluated as a number of possible assessment outcomes between the calculated threshold \( T \) value and the highest possible score. If the number is less than a predefined value, the assessment is also rejected.
7. If the assessment was rejected, the generation procedure starts again.

This procedure can be also used to evaluate the reliability of an assessment. For that step 2 should be excluded from the generation procedure, the other steps are the same.

An online assessment web application based on the proposed method has been developed in the form of a python web application.

It uses JSON files as data storage.

The time for each assessment is limited.

The assessment for each student is unique in terms of the uniqueness of the task set in an assessment. Some of the tasks can be presented in more than one assessment, but the whole combination of the tasks for each assessment is unique. The more tasks we have in the task pool, the fewer common tasks each particular assessment shares with the other ones.

For each assessment, its own passing threshold is calculated according to the procedure described above. The information about each attempt is stored as a JSON file, including the date and time of the assessment creation, the time when the assessment was started and finished, all stems, all options, a list of answers and distractors, the testee’s response, the calculated passing threshold, and the distribution of the calculated outcomes.

The assessment process is shown below in the form of a sequence diagram (Fig. 5):
The application has been used for three years for assessing students of Siberian State University of Science and Technologies. A comparison between the proposed methodology results and the classic exam shows the high efficiency of the proposed approach.

7 Actual Outcomes

The proposed approach is a convenient means allowing to assess knowledge, skills, and abilities of students. It provides an objective way to generate an assessment, allowing us to evaluate its discrimination ability and its testees’ ranging ability. The discrimination and ranging abilities of assessments generated by the proposed procedure highly correlate with the ones for the traditional exams.

The on-the-fly generation of the assessment allows us to provide each testee with a unique assessment, preventing the assessment cheating by decreasing the possibility to discuss the tasks of the assessment the testee passes at the moment among the other testees or third parties.

The implementation of the system in the form of online assessment system also allows us to use the information about testee’s answering process to prevent cheating using machine learning technologies.

To overcome the proposed approach it would require consolidated testees’ efforts to collect and to answer all the tasks in the task pool. If the time window for the testing is limited, this possibility of cheating the system is also reduced.

8 Conclusions

The proposed approach can be easily integrated into an existing teaching process, especially if the online assessment systems are in use already. Using the described approach allows us to increase the objectiveness of an assessment process. The
online assessment system can be used both for scheduled surveys and for pre-exam surveys, as well as students’ exams self-preparing.

Surveys performed in the university show that the differentiation ability of the presented approach is at least at the same level comparing with more traditional approaches like colloquiums and interviews, providing the tasks of the assessment are built correctly.

As a drawback of the described approach, the complicated procedure of the assessment threshold calculation method can be mentioned. Sometimes it’s hard to explain to testees how their results have been got. The threshold calculation procedure itself also takes a noticeable time as it requires the continuously repeating answering process simulation.

One should note that the multiple-choice assessment, as well as multiple response questions, are not a substitution for traditional exams or interviews, especially because of different abilities of students in the field of communication and social interaction.

References

1. Ali, S.H., Carr, P.A., Ruit, K.G.: Validity and reliability of scores obtained on multiple-choice questions: why functioning distractors matter. J. Scholarsh. Teach. Learn. 16(1), 1–14 (2016)
2. Aguaded-Gómez, J.I.: The MOOC revolution: a new form of education from the technological paradigm. Comunicar 41(21), 7–8 (2013)
3. De Champlain, A.F.: A primer on classical test theory and item response theory for assessments in medical education. Med. Educ. 44(1), 109–117 (2010). https://doi.org/10.1111/j.1365-2923.2009.03425.x
4. Downing, S.M.: Reliability: on the reproducibility of assessment data. Med. Educ. 38(9), 1006–1012 (2004). https://doi.org/10.1111/j.1365-2929.2004.01932.x
5. Forehand, M.: Bloom’s taxonomy. Emerg. Perspect. Learn. Teach. Technol. 41(4), 47–56 (2010)
6. Cornock, M.: Scaling up online learning during the coronavirus (Covid-19) pandemic (2020)
7. Kehoe, J.: Writing multiple-choice test items. Pract. Assess. Res. Eval. 4(1), 9 (1994)
8. Moreno, R., Martínez, R.J., Muñiz, J.: New guidelines for developing multiple-choice items. Methodology 2(2), 65–72 (2006)
9. Ng’ambi, D., Bozalek, V.: Editorial: Massive open online courses (MOOCs): disrupting teaching and learning practices in higher education. Br. J. Educ. Technol. 46(3), 451–454 (2015)
10. Sangaiah, A.K. (ed.): Deep Learning and Parallel Computing Environment for Bioengineering Systems. Academic Press, Cambridge (2019)
11. Towns, M.H.: Guide to developing high-quality, reliable, and valid multiple-choice assessments. J. Chem. Educ. 91(9), 1426–1431 (2014)
12. Cook, D.A., Beckman, T.J.: Current concepts in validity and reliability for psychometric instruments: theory and application. Am. J. Med. 119(2), 166–e7 (2006)