Multiple Choice Assessments: Evaluation of Quality

Alexander Sayapin
Applied Mathematics Chair
SibSAU
Krasnpyarsk, Russia
alstutor@gmail.com

***Abstract*— Multiple choice assessments are widely used in modern higher education. A lot of different best practices that describe how to build a good assessment are known, but there is no way to evaluate how good the assessment is, or how difficult is it, or how many stages of competencies it can unveil. The method to evaluate the difficulty and the differentiate ability of a multiple choice assessment is described in this article.**

Keywords— choice assestment, complexity, differentiation ability, statistical approach, simulation

I. INTRODUCTION

A lot of works are devoted to the problem how to build a good multiple-choice (MC) assessment [1, 2, 3]. Most of them describe the methods to formulate stems and responses. But the practice tells us that even if the MC test is well-formulated and valid, there still can be situations when teachers are not satisfied with the results, no matter are the results too bad or too good.

Nowadays there are two main approaches to the process of knowledge, skills and experiences evaluation:

"Norm-Referenced Assessment: A test or other type of assessment designed to provide a measure of performance that is interpretable in terms of an individual's relative standing in some known group. Criterion-Referenced Assessment: A test or other type of assessment designed to provide a measure of performance that is interpretable in terms of a clearly defined and delimited domain of learning tasks." [4]. In other words, the main difference is the method of definition of the threshold level. In the first case we mean that, for example, 90% of students should be considered as successfully passed the test, no matter how strong their knowledge and skills are. In the second case we set the predefined threshold level. Only if the student's knowledge and skills are above the level, we consider him or her as one who successfully passed the test.

Corresponding to the Bologna Process, even at the first cycle of qualification the students should demonstrate their ability to apply their knowledge and understanding for solving problems within their field of study [5]. It means that criterion-referenced testing (CRT) is much more preferable in modern higher education.

Today CRT is the mainstream of the educational assessment. This method is easy to use, it is well-formalized, it takes a few of time.

II. TEST FORMATS

All the multiple-choice test items can be divided into two groups: true/false-items and one-best-answer items [6]. There are some benefits and drawbacks of each method. The main criticism of the true/false items is that it's hard to decide if the answers are absolutely correct or incorrect. One-best-answer type of items is more preferable [6]. One should note that problems with true/false items are connected with the method of counting of the right answers percentage. We will discuss this problem later.

Also, it is much more preferable to give the unique variant of the test to each student. It lowers the risk of a situation when students help each other to pass the test, what can occur sometimes.

To avoid such a situation, we can form the big pool of the items, and to form the assessment for the given student we can choose some set of the items. For another student the set will be completely or partially different.

III. METHOD FOR FAIR ANSWER'S EVALUATION

Let us consider some extraction from the multiple-choice test, developed for the artificial intelligence course. There are only 3 items in this example:

Definition of the intelligent term is based on

1. environment modeling

2. action planning

3. intelligent system structure

The system that uses the set of connected elements that models activity of the parts of a rat brain is based on

1. bionic approach

2. conceptual approach

3. heuristic approach

System that is developed for image recognition using psychological theories can be described as

1. system based on bionic approach

2. system based on conceptual approach

3. system based on heuristic approach

The first item belongs to the true-false family. It has two answers (first and second), and one distracror (third).

The second item as well as the third belongs to the true-false type too; however it has only one answer (first), and two distractors. The third item also has only one answer (second) and two distractors.

Discussing the method to check the correctness of the students choice, we can see two main approaches to evaluate the student's knowledge: consider the answer of the student as wrong is he or she made even one mistake (for instance, didn't mark one answer of two in one item, or checked one distractor while the other elements are absolutely correct), and calculate some measure of correctness of student's answers. From the point of view of a student, the second approach is fairer, and even the teacher that uses the assessment may agree with that.

Now we offer the method to calculate the measure of the correctness of answers that the student gave.

To calculate the measure of the correctness, we can use very simple approach: just calculate how many answers the student marked, and then divide it on the whole number of answers. For instance, let's consider the situation when the student marked variants 1 for the first item, 1 for the second and 1 for the third. Let me remember that for the first item the answers are 1 and 2, for the second one the answer is 1 and for the third one answer is 2. So we calculate the number of answers the student marked . It is 2 (one of two for the first item, one of one for the second and zero of one for the third). Now we should divide 2 on 4 (the whole number of answers in the assessment). So we get 0.5, or 50% of correct responses given by the student.

However, the calculation of the correctness of the student's responses may lead to false positive reaction: if the student marks more variants than necessary for every item (all the variants for example), he or she gets high mark for his or her knowledge. To prevent such a situation we should use more complicated approach, and take into account not only answers, but distractors too.

The most appropriate way is to use the following formula (1)

$$R = \min\left(\frac{M_s}{M}, \frac{U_s}{U}\right), \qquad (1)$$

where R – level of student's knowledge

M – number of the answers in the whole test,

U – number of the distractors in the whole test,

M_s – number of the answers marked by the student,

U_s – number of the distractors not marked by the student.

It allows us to take into account even the partial knowledge of the student.

Another important question is to determine, how high the score should be to pass the test? Should it be 50%, or 60%, or even 75%? This important question we will consider later.

As to one-best-answer items family, the approach to estimate the correctness of the student's answer is different. We can calculate correctness of the student's answer the same way as one usually did it for the true-false items: calculate the percentage of the correct answers chosen by the student, but it's slightly inappropriate too.

The more preferable way is to range all the answers and distractors due to their correctness and give them a value of some coefficient that is in range of 0 to 1. So we may use the average value of the coefficients for all the items as the measure of student's knowledge. But the problem with the threshold level still appears in this case.

How we can determine the threshold level?

IV. How to determine the threshold level

It is obvious that process of passing of the test is a process that we can describe statistically. In general case the student does not know all the answers so he cannot give all correct responses. The question is which number of right responses is enough to pass the test?

To answer this question we can use some statistical approach, simulating the answers of the student that answers our questionnaire by random. Performing a lot of such attempts, where the computer answers the test items randomly, we finally get a statistic, some kind of distribution (fig. 1).

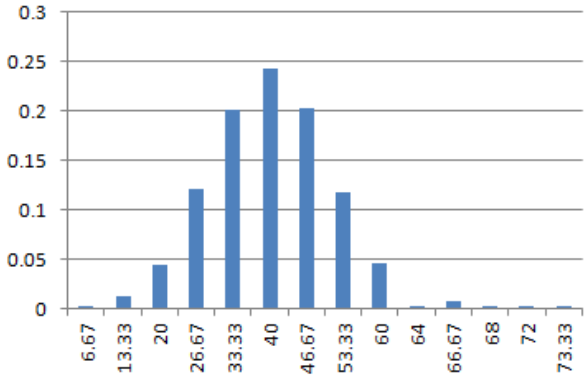

Fig.1. The distribution of the percentage of correct responses. $\alpha + \beta = \chi.$ (1) (1)

The answer to the question if the student successfully passed the test is equivalent to checking of statistical hypothesis H_0: the percentage of the right responses given by the student is more than the one we can get by random with the given probability.

We set the probability α (for example, α=0.05, it means that in one case of twenty the student that answers by random can pass the test successfully), and then try to find the threshold level *Rs*, which probability is higher than value we set for probability. For the fig. 1 this value is equal to 53.33.

One of important questions is "When to stop simulation? How much tries is enough?". We can answer this question

analyzing how the percentage of the correct responses changes. The more tries we do, the less percentage changes. So we can set the level of the changes of percentage, for example, equals to 0.01%. When the percentage of correctness of two consequent tries differs less than 0.01%, we can stop simulation.

The threshold level allows us both to determine did the student pass the test successfully or not, and to determine the score of the student.

Let the score be in the interval of 0 to 100%. If the student get the *R* value that is less than *Rs*, his or her score is equal to 0. If the *R* value is greater than *Rs*, we should normalize the difference between *R* and *Rs* so it should be between 0 and 100%. To do this we can use formula (2):

$$S = (R - R_s) \cdot \frac{100}{(100 - R_s)}, \qquad (2)$$

where *S* – score of the student

R – level of student's knowledge,

R_s – threshold level.

We use this formula providing *R* value is more than *Rs* value, otherwise S value is equal to 0.

Sometimes a problem, connected with misunderstanding of values S, appears. It is necessary to explain students what 0 means for this score, that 0 means that the knowledge level of the student is indistinguishable from random answers.

Now let's discuss the multiple-choice assessment's properties.

V. Multiple-choice test properties

In a connection with a multiple-choice test the validity, reliability, difficulty, differentiate ability terms are often used. In some cases these terms relates to the test items, and sometimes they are used in application to a test itself.

Most of the terms can be determined using expert evaluation. The expert determine how valid the test (or test item) is, how difficult it is, and so on. But all these evaluations are subjective. If we invite two experts, it is more than possible that they evaluate the same test (or the same question) differently.

The difficulty of the test traditionally refers to the ratio of the students that successfully passed the test to the number of students that didn't. But it means that difficulty is not an objective test's property and its value depends on students we test. To evaluate the test instead of students we should use another characteristic.

The ability to differentiate the students that passed the test and who didn't, or to grade the students that passed the test due to their knowledge and skills we could call the differentiation ability. In most cases we cannot measure this ability without some using of the test.

In this work we discuss only difficulty and differentiate ability of a test. The validity and reliability of a test are still too complex properties to evaluate them without expert's opinion.

Let us consider the way to evaluate the difficulty and the differentiation ability of a test objectively.

VI. Calculation of multiple-choice assessment difficulty

Due to formula (1), the student passes the test successfully only providing he or she gets *R* value that is more than *Rs* value.

There are two definitions for the test difficulty: the first one refers to the percentage of students that successfully pass the test, and the other shows us how easily the student can get higher mark for his or her responses. The first approach, as it's already mentioned, cannot be considered as an objective approach.

To implement the second one we may use the threshold level *Rs*. The higher threshold level is the more complicated the test is. Also we may consider Rs value as the measure of the objective test's difficulty with given value of probability.

Another important property of the test is differentiating ability. One may consider this property as number of levels to which one can attribute the given student due to his or her answers.

This number of levels cam be determined by simulation of student's responses, as it is described earlier in the section iiv.

According to the fig. 1, the number of levels of the test is equal to 14. To differentiate the students that successfully passed the test we may use 7 level, which values are equal or higher than 53.33%.

Summing up, we can define the difficulty of the test and its differentiation ability as follows:

the difficulty of the test is the threshold level *Rs*, calculated as the percentage of the correct responses that can be given randomly with the given probability;

the differentiation ability of the test is the number of percentage levels of correct responses that can be given randomly.

One should note that this kind of difficulty of the test isn't the "real" one. It doesn't show how many students of the given group will successfully pass the test, it only shows us how hard is to get higher score for the test.

One more important thing is that both difficulty of the test and differentiation ability of the test should be applied to each test separately. So, if one forms the test that consist of the items taken from common pool of items, both properties should be calculated for test, not for pool of items, and not for each item too.

VII. PROGRAM IMPLEMENTATION

While developing the program for this approach we tried to implement a few main ideas:

versatility;

interoperability;

integrity.

In application to our task, versatility means that we should be able to use this system to taking the test for practically any subject or course or discipline. Taking into account the fact that test may be performed in different ways, such as internet/intranet testing and in printed form; we should develop the system, taking any kind of responses.

Interoperability means that we should be able to use the system on any platform, at least on three most popular platforms such as Microsoft Windows, OS X and Linux.

Integrity means that we should be able to form any kind of a test (in printed or network form) from the same pool of items, that can be used on any platform.

Due to these requisites we have to choose the way to describe multiple-choice assessment items (e.g. file/database format), the programming language and the server part of the system.

As storage method the XML file format, the JSON file format and the MySQL database were considered. Each one has its own benefits and drawbacks. For example, MySQL database is very popular storage method and it is fast enough to get or put data into the storage, but discussing the interoperability we have to mention that we cannot use the same instance of the database engine on Microsoft Windows and Linux. The simple file formats are more preferable.

The JSON file format is more modern format, it has smaller footprint, it is more friendly for human reading and modifying, but there are a lot of different program libraries and frameworks for XML format. So we decided to use XML.

As to program language, we considered a few contenders: Java, C#, C/C++, Python, PHP. All these languages are portable and can be used on all the platforms we considered. All the languages are object-oriented (except PHP and C). But a program written in C/C++ cannot be used on any platform directly, it has to be recompiled. Python and PHP don't implement the static typing paradigm, so the process of debugging may take a lot of time.

So as the primary languages Java and C# were chosen. Now the system is implemented in C#. The main reason is its possibility to run on any system that has runtime engine implementation (for example, .Net Framework on Microsoft Windows and Mono on Linux and OS X). This language also has a lot of different libraries and possibilities to work with XML, internet/intranet and so on.

To achieve the versatility, the main library was developed as an object library. It describes a lot of universal and specific objects correspond to test, assessment, item, stem, answer, distractor, student and so on.

It allows to use this library to develop, maintain and check the test in both printed and internet/intranet forms. Now two main applications are developed. The first one is developed to check the printed form of the test assessment. This program allows also to prepare PDF file with the on the basis of assessment XML file. The system prepares the particular test for each student.

VIII. CONCLUSION

The approach described in this article is used in Siberian Aerospace State University for student's knowledge and skills evaluation. The score of the student, estimated by teacher corresponds to the ones in such kind of assessments.

Now we are going to improve the approach so it allows to form test of higher quality in terms of test's difficulty and differentiation ability evaluation.

REFERENCES

[1] Steven J. Burton, Richard R. Sudweeks, Paul F. Merrill, Bud Wood Multiple-Choice Test Items: Guidelines for University Faculty, Brigham Young University Testing Services and The Department of Instructional Science, 1991.

[2] Derek Cheung, Robert Bucat How can we construct good multiple-choice items? Science and Technology Education Conference, Hong Kong, June 20-21, 2002.

[3] Berk, R.A. Criterion-Referenced Measurement. Baltimore, MD: John Hopkins Press, 1980.

[4] Linn, R. L., & Gronlund, N. E. (2000). Measurement and assessment in teaching (8th ed.). Upper Saddle River, NJ: Prentice Hall

[5] The framework of qualifications for the European Higher Education Area. http://www.bologna-bergen2005.no/EN/BASIC/050520_Framework_qualifications.pdf

[6] Susan M. Case, PhD and David B. Swanson, PhD Constructing Written Test Questions For the Basic and Clinical Sciences (3thed.). National Board of Medical Examiners, 2002.