Available online at www.sciencedirect.com
ScienceDirect
Computers & Education 51 (2008) 743–756
COMPUTERS & EDUCATION
www.elsevier.com/locate/compedu

Some problems of computer-aided testing and "interview-like tests"

Denis V. Smoline *

Altai State Technical University, Information Technologies in Economy, Department of IT, Lenin Avenue 46, 656019 Barnaul, Altai, Russian Federation

Received 16 April 2007; received in revised form 23 June 2007; accepted 26 July 2007

Abstract

Computer-based testing – is an effective teacher's tool, intended to optimize course goals and assessment techniques in a comparatively short time. However, this is accomplished only if we deal with high-quality tests. It is strange, but despite the 100-year history of Testing Theory (see, Anastasi, A., Urbina, S. (1997). *Psychological testing*. Upper Saddle River, NJ: Prentice-Hall) there still exist some misconceptions. Modern wide-spread systems for computer based course management and testing reveal a set of problems corresponding to certain features of testing methods.

This article is devoted to some omissions typical to several course management systems (e.g., Moodle and Blackboard). These omissions and the ways of avoiding them are shown in a simple test intended to verify student knowledge. We suggest a special test description language dedicated to drawing your attention to the mathematical aspects of test quality. The language can also be realized in computer software. We provide an example of such software in this article.

Keywords: Computer-based testing; Teaching techniques; Course management systems; Moodle; Blackboard; Chopin; Test description language; Test validity; Economy; Reliability; Difficulty of a test; Guessing probability; Measures of similarity; Inseparable hypotheses

1. Introduction

The central point of the classical Testing Theory (see Anastasi & Urbina, 1997) has to do with test quality. The quality of a test (or a part of a test) can be evaluated with the help of certain statistical methods. The basic parameters of test quality are: test validity, test reliability and test economy. In general they can be defined as follows:

- *test validity* – ability of the test to measure the target characteristic;
- *test reliability* – ability of the test to remain valid over a series of executions – either in several groups of students or several series of the same test in the same group.

* Corresponding author. Tel.: +7 3852 482005.
E-mail address: sdv56@mail.ru

0360-1315/$ - see front matter
doi:10.1016/j.compedu.2007.07.008

- *test economy* – ability of the test to be valid and reliable with a minimum of expense. Depending on the aims of the testing, the expense could be represented in terms of the number of questions in the test, total time of testing, the amount of ‘‘mental loading’’ per student, and visual fatigue[1] (e.g. see Mourant, Lakshmanan, & Chantadisai, 1981; Nakazawa, Okubo, & Suwazono, 2002) etc. On the other hand, test economy implies the taking of the maximum of information about the student’s characteristics from the available testing data.

Another important point of Testing Theory relates to test content. Here are two correlated parameters:

- the difficulty of a test (or a part of a test). In practice, the difficulty of a test is an expert mark formulated in terms of fuzzy logic, e.g., ‘‘for beginners’’, ‘‘intermediate’’, ‘‘advanced’’, etc. The difficulty of a single test item is usually indicated by a real coefficient (in terms of Zadeh logic). A coefficient of 0.7. for example, means that the test item is really difficult.
- guessing probability – the probability of getting a high score (for the entire test or a test item) without any knowledge of the subject matter being tested. In other words a very difficult test item can be created using a testing method that enables the answer to be easily guessed by the student. From this point of view, ‘‘True/False’’ tests are rather unreliable. ‘‘Fill the Blank’’ tests, on the other hand, are quite effective. In fact, however, it is best not to create all of the test items as ‘‘Fill the Blank’’ type questions. The validity of such a test is low, however, the validity of ‘‘True/False’’ test item is relatively high. Sometimes students cannot understand a ‘‘Fill the Blank’’ test due to inaccuracy or polysemy.[2] This problem is typical of good students or adult students who are experts in the subject of the test. At first sight, it seems that this problem could be solved by the detailed formulation of every test item. But there is a flip side of such a solution in terms of increased test item length, total time for testing,’’ mental load,’’ etc. Still, this is not the worst consequence! Experts doubt that the question could really be so simple, so they try to find an underlying trick. It is curious to note that while experts’ test results are often very low, when interviewed, these students show high competence in the subject matter. This is a case of the so-called ‘‘face validity problem’’ (problem 0). It mostly depends on the type of test item and variants used. To evaluate test validity it is convenient to use an interview method (e.g., see Hamilton, Nussbaum, & Snow, 1997). Another way is to make the test adaptive. However, it is necessary to follow a set of rules, otherwise the test will be non valid, inaccurate and redundant. (e.g., see Häusler, 2006).

Accordingly, we believe that the test should structurally consist of items having various levels of difficulty. If we take the test final score as the percentage of correct answers (the formula used in Moodle and Blackboard), than we face a risk of:

- giving a high score to a student who guessed the correct answers (problem 1),
- giving a low score to a good student who knew the correct answers but found the test non-valid (problem 2).

The above mentioned problems can be mitigated by applying to the final score a non-linear formula calculated on the basis of an algorithm assigned by the test author.[3] If this is done, the test will become adaptive and ‘‘interview-like’’.

[1] For e.g., it is essential in some psychology tests.

[2] The problem may be solved by the use of the so-called ‘‘measures of similarity’’, such as Hamming distance, Tanimoto measure or Levenstein distance (see Kohonen, 2001). These measures are used in ‘‘Intelligent Edit (IEdit)’’ operator (see Table 4).

[3] To use such an algorithm the conditional jump and local variables should be included in the test structure. In this case, the test will consist of several groups of test items with the conditional jump between them. Each group will allow the teacher to prove a hypothesis of the test taker. In Moodle and Blackboard these elements are not supported: the more tasks (or cycles of testing) the test has, the more exact the measurement of knowledge is. This is not true, because it is necessary to analyze not only the percentage of the correct answers, but also the sequence of the correct and incorrect answers.