Dynamic Reliability Assessment of Multiple Choice Tests in distributed E-Leaning Environments: A Case Study

Ioanna Likourentzou, George Mpardis ,Vassilis Nikolopoulos and Vassili Loumos
National Technical University of Athens
School of Electrical and Computer Engineering,
Iroon Politechniou 7, Zografou, Athens, GR 15773
{ioanna,gmpardis,vnikolop}@medialab.ntua.gr, loumos@cs.ntua.gr

Abstract. The development of high-quality e-learning products is one of the most demanding areas in the field of educational research. Reliability of the students' grading mechanisms especially in the case of virtual classrooms, which lack in physical student-instructor interaction, is extremely important. In this paper, based on real data, we utilize two reliability estimation methods to calculate several multiple choice tests' reliability. Moreover, since multiple choice tests are an imperfect measure of students' knowledge, we also estimate the students' true ability of scoring using the tests' standard error of measurement. Concluding this study embeds reliability assessment methods in the e-learning process and then carefully analyzes the produced data to provide the strengths and weaknesses of the analyzed course's multiple choice tests.

1 Introduction

An issue that nowadays concerns both the educational industry and the research community is the subject of e-learning evaluation. As e-learning includes more technological than human factors, compared to conventional education, its proper evaluation is vital. In an e-learning environment the trainer does not have direct contact with the trainees and this may result in a difficulty of proper student grading. The only student evaluation means is via assigned open-answer projects and multiple choice tests that the students deliver. Subsequently it is obvious that e-learning courses aiming to provide trainees with high quality education should consist of highly reliable projects and multiple choice tests which will accurate measure the students' level of knowledge. Especially in the case of multiple choice tests, which

are not subjectively corrected by a human trainer, the reliability issue is even more significant.
In this paper we estimate the reliability of several e-learning multiple choice tests, the error of measurement they present, compare them and then discuss the results.

2 Educational Framework

The learning object of this e-lesson is an introductory course to computer network communication. The course runs in the open source platform Moodle. The existing analyzed data were derived from three classes corresponding to three semesters (Fall 05, Spring 06 and Spring 07) of approximately ten weeks each. Every class took four twenty-question multiple choice tests, each test corresponding to four learning weeks in the duration of the e-learning course. The students could choose to repeat the test in order to achieve better results, therefore providing our study with the necessary data to use with the test-retest method. This is considered a rare opportunity not only in the e-learning field but generally in education, because usually students do not take the same test more than once.

3 Methods of calculation and Standard Error of Measurement

3.1 Multiple choice test reliability

A multiple choice test reliability is the extend to which this test produces consistent, stable, trustworthy and repeatable results when administered by the same group of students twice [1]. For a particular set of test/retest scores one can plot the scores on a scatter-gram to obtain a general idea of the reliability that this test presents. The more the sets of scores deviate from the x=y line the more unreliable they are.

3.2 Methods of calculation

There are two methods to estimate a multiple choice test's reliability based on the number of times this test was administered by the same group of students:
a. Multiple-administration methods require two or more assessments of the same test. The most appropriate multiple-administration technique for multiple choice tests is the test/retest evaluation method [2], which lies in having the same group of students take the same test two times. If the test is reliable most of the examinees will tend to get the same or very similar scores on both administrations. The evaluator can use a coefficient to calculate the test's reliability. The most widely used one the Pearson product-moment correlation coefficient (PMCC) between two administrations of the same measure.
Assuming that X and Y are the two sets of student scores, then the test's reliability based on PMCC is defined as following:

$$r_{tr} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}$$

b. Single-administration methods are used in cases when a test can be assessed only once [3]. A typical single-administration method is the split-half reliability technique in which the evaluator divides the test into two halves and treats them as alternate administrations of the same test. The reliability correlation of these two halves is then calculated. Since this correlation is based only on half the test length, it cannot be fully indicative of the tests' reliability. To fix this inaccuracy we used the Spearman-Brown prediction formula which predicts the full-test reliability based on the half-test reliability correlation. The Spearman-Brown split-half reliability is defined as following:

$$r_{SB} = \frac{m^* r_{XY}}{1 + (m-1)^* r_{XY}}$$

where

r_{SB} is the Spearman-Brown split-half reliability
r_{XY} is the Pearson correlation between forms x and y
m is the total sample size divided by sample size per form (m is usually 2)

As with other split-halves measures, the Spearman-Brown reliability coefficient is highly influenced by alternative methods of sorting items into the two forms. We used a random assignment of items to the two forms as this is considered to be amongst the most effective means to assure equality of variances between the forms.
The above methods of reliability estimation should not be expected to be equal since they are prone to different sources of error.
Due to the type of our available data (some of the examinees had the opportunity to take the multiple choice tests twice in order to optimize their scores while others took the tests only once), in this study we will be able to use both multiple and single administration methods .

3.3 Standard error of measurement

Multiple choice tests are an imperfect measure of a student's knowledge level since they may be influenced by extraneous factors such as chance error, differential testing conditions, imperfect reliability and other errors of measurement. A way to estimate a band or interval within which a person's true score (true ability of the student) would fall is the standard error of measurement (SEM). SEM is calculated using the standard deviation and the reliability of test scores and represents the amount of variance in a score resulting from factors other than achievement. [4]:
The SEM is calculated using the formula:

$$\text{SEM} = \sigma_x \sqrt{(1 - r)}$$

where σ_x is the test's standard deviation and r is the test's reliability estimate.

3.4 Reliability coefficient interpretations

Table 1 indicates the evaluation of a reliability test/retest correlation coefficient r_{tr}. Note that as the coefficient's values decrease so the proportion of the incorrectly awarded examinees increases [5]:

Table 1 Reliability Interpretation

Reliability (r_{tr})	Coefficient Evaluation
0.90 – 1.00	High reliability - Appropriate for the assessment of a student on the basis of a single test score.
0.80 – 0.89	Acceptable reliability. Appropriate for the evaluation of an individual student if averaged with a few other scores of similar reliability.
0.60 – 0.79	Low to moderate reliability. Appropriate for the evaluation of a student only if averaged with numerous other scores of similar reliability.
0.40 – 0.59	Uncertain reliability. Should be used with great watchfulness when evaluating individual students. May be suitable for the calculation of average score variations between groups

Values of 0.80 and higher, are generally considered to be satisfying. However, one should not be based on a single test score to make significant decisions about individual examinees when the corresponding reliability coefficient is less than 0.80.

4 Results - Discussion

4.1 Comparison of the scores on the same test on different semesters

The mean of the grades that the students accomplished in all three semesters, as depicted in figure 1, seem to be consistent. This means that all three classes received approximately the same scoring results. If the trainer is based only on this observation, the tests all seem to be appropriate for student evaluation. However, following it is analyzed that based on their reliability the tests do not prove to be equally appropriate for an accurate and fair student evaluation.

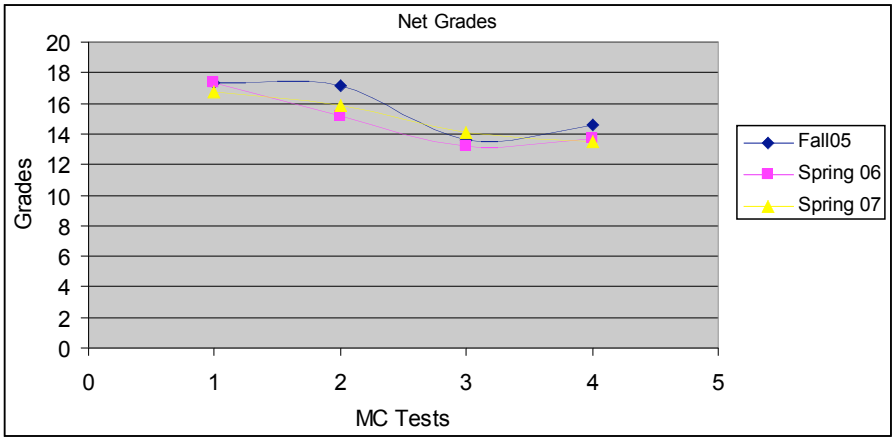

Fig. 1. Multiple Choice Test Grades

4.2 Tables with calculations of reliability for each lesson/semester

4.2.1 Spearman-Brown Reliability Results

The Spearman-Brown method indicates (figure 2) that the first two multiple choice tests resulted in the most reliability inconsistencies. MC1 yields from near-zero (0, 11) to low (0, 62) reliability and MC2 yields from uncertain (0, 46) to acceptable (0, 85) reliability. These reliability results indicate that MC1 and MC2 tests (and especially MC1 that did not achieve acceptable reliability in neither of the three course classes) need to be ameliorated in order to be solely trusted for student evaluation. At this point MC1 should only be utilized to assess students' comprehension only if supported by other assessment sources such as written projects, forum participation and correctness of answers in teachers' on-line questions. The above are also applied to MC2 with the exception of class spring '06 for which the test scores yielded acceptable reliability. Nevertheless, our suggestion regarding this test is to be treated with carefulness since the high reliability value of class Spring 06 may be due to the test scores' splitting. Tests MC3 and MC4 produce moderate (0, 71) to high reliability (~0, 9) and are thus considered appropriate for student assessment. Consequently, alternative assessment sources like the ones mentioned above are auxiliary but not necessary for the evaluation of the students of these learning weeks.

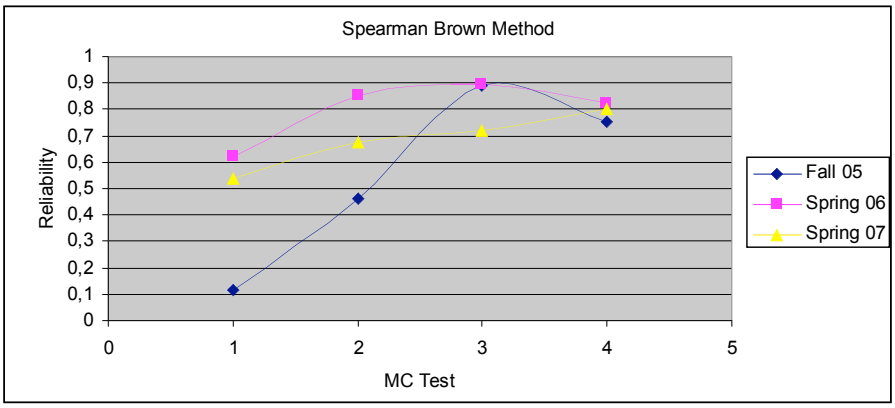

Fig. 2. The multiple choice test reliabilities calculated according to the Spearman-Brown prophecy formula.

4.2.2 Test – Retest Reliability Results The test retest method is generally expected to produce lower reliability results that the spearman brown prophecy formula. This is explained mainly due to the memory effect that influences the students in their scoring. The above means that, since the students were allowed to retake the test in the same learning week they tend to remember some of the questions and thus score better. In most cases their scoring increases compared to the first test-taking, thus lowering reliability. This is mostly expected to occur in the latter multiple choice tests and not in the first one mainly due to the fact that the students need some time to be accustomed to the e-learning environment and take advantage of the memory effect on retaking the test. According to the test retest method, the first two multiple choice tests yield the most reliability inconsistencies. MC1 yields acceptable reliability for classes Spring 06 and Spring 07 while it yields an uncertain reliability for class Fall 05. The second multiple choice test MC2 generally yields very low reliability for classes Fall 05 and Spring 07 with the exception of class Spring 06 where it yields acceptable reliability. On the other hand multiple choice tests MC3 and MC4 although they yield low to moderate reliability they tend to produce these results consistently. Due to the memory effect and also the fact that less students participate in the test re-taking, the test-retest method is less indicative of the true tests' reliability.

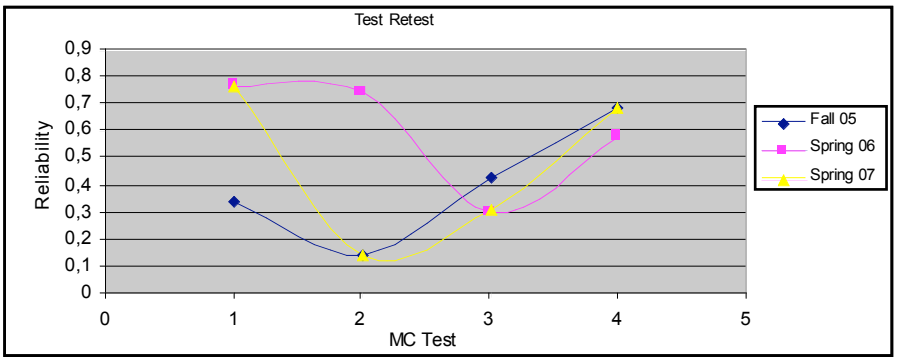

Fig. 3. The multiple choice test reliabilities calculated according to the test-retest method.

4.3 Student standard error of measurement As mentioned above, the tests are not perfectly reliable and thus, a student's observed score and true score will differ. The standard errors of measurement that yield the range of values that would most likely contain the student's true scores are depicted in Table 2. The more unreliable the test scores are the more standard error of measurement these tests yield.

Table 2 Student scores' standard error of measurement

	Fall 05		Spring 06		Spring 07	
	Reliability	SEM	Reliability	SEM	Reliability	SEM
MC1	0,11	2,02	0,62	1,31	0,54	1,60
MC2	0,46	0,91	0,85	1,62	0,67	1,79
MC3	0,89	1,26	0,90	1,30	0,72	1,89
MC4	0,75	1,40	0,83	1,70	0,80	1,60

Following we provide an example of the evaluation process that the instructor could perform based on the previous results. The example utilizes a randomly selected student and can be applied to any course student. Figure 4, depicts the standard error of measurement on the grades of a student that belongs to the class Fall 05. MC1 yields the highest error of measurement, followed by MC4, MC3 and MC2. Moreover, MC1 produced a very low reliability and this fact along with the high SEM it produces, should make the instructor very careful on utilizing this item in the student evaluation process. MC3 yields both low SEM and high reliability and is thus the most indicative of the students' progress. MC2 and MC4 should also be taken into consideration in the evaluation process, but especially MC2, although it produces the lowest SEM should not be highly valued since its reliability is uncertain.

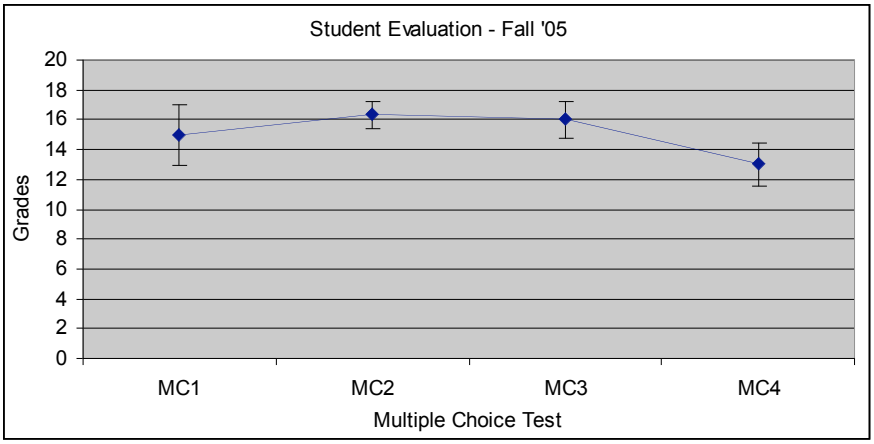

Fig. 4. Scores and SEM for a random student – Fall 05

5 Conclusion

This study focuses on e-learning reliability by estimating the reliability of several multiple choice tests of an introductory e-learning course. The tests' distinctive accuracy is precisely calculated using reliability methods (Spearman-Brown formula and the test-retest method) and the standard error of measurement (SEM), thus enabling trainers to improve their evaluating process. This methodology significantly contributes in delivering a better and more reliable e-learning course by making it strongly competitive and trustworthy.

References

1. Wallace R. Blischke, D. N. Prabhakar Murthy ,Reliability Modeling Prediction and Optimization , Wiley Series in Probability and statistics,2000, p.18,19
2. Renee Bradley, Reliability Issues And Evidence, U.S Department of Education, Office of Special Education Programs (OSEP), Toolkit on Teaching and Assessing students with Disabilities
3. Richard A. Zeller, Edward G. Carmines, Reliability and Validity Assessment, Sage Publications, 1979
4. Rudner, L. M., & Schafer, W. D,Reliability, (ERIC Digest). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. (ERIC Document Reproduction Service No. ED458213) ,2001
5. Leo M. Harvill, An NCME Instructional Module on. Standard Error of Measurement, Educational Measurement: Issues and Practice 10 (2), 33–41, 1991