Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej Nr 37

I Konferencja

e-Technologies in Engineering Education eTEE'2014

Politechnika Gdańska, 30 kwietnia 2014

THE INFLUENCE OF QUESTION SET ON STUDENT QUIZ RESULTS

Artur OPALIŃSKI[1]

1. Gdańsk University of Technology, Narutowicza 11/12, 80-233 Gdańsk, Poland
tel.: +48 58 347 13 32 e-mail: Artur.Opalinski@pg.gda.pl

Abstract: The advent of e-Learning tools allowing for automated online test grading will probably increase the frequency of using tests in technical education. The same tools may provide for measures of test question quality. By purposely crafting question sets, test grading may serve different goals. The paper contains examples and test study with score histograms.

Keywords: test question quality, test assessments, e-Learning

1. INTRODUCTION

Multiple choice test questions are considered a convenient way of checking educational achievements, for they offer easy marking and easy evaluation of question quality [1]. The latter is less known in Polish education, especially in the technical education which did not yet embraced test questions [2].

Application of e-Learning tools makes test questions even a more convenient approach, because marking students results and evaluating of question quality can be done automatically. This makes test questions particularly appealing for assessments in massive online courses. Such courses begin to appear in Polish academia, as cost cutting measures results in creating bigger students groupings. While the number of students enrolled to a single course is far from the new educational phenomenon called MOOCs [3], they happen to reach already hundreds of students, which makes non-automated grading techniques hard to imagine.

The e-Learning tools, for example the free Moodle platform [4], offer additional benefits to traditional tests:

- managing question sets in question database,
- random selection of questions per each test attempt and per each student,
- randomization of answer options order,
- automated grading,
- automated evaluation of test questions quality,
- sharing individual test scores with the corresponding students,
- presenting students scores individually, and in the form of histogram.

Random selection of questions prom the set, together with randomization of answer options for multiple choice test questions provide quite strong protection against cheating between students during the test.

To provide for full protection against dishonesty, tests should be run in class, under strict teacher supervision, as they traditionally used to be run. This is also important to protect against a new threat which arises with the advent of online tests – against stealing of the questions from the database. Copying of question content in a typical Web browser is plausible, and generally easier to do on a massive scale than copying of questions distributed on paper sheets. Of course, having a big questions database diminishes the problem of stolen questions, but creating a big database of good questions is a challenging task.

The rest of this paper is devoted to the problems of ensuring good quality of test questions and putting them into tests to create effective assessments.

2. PROBLEM

With that many advantages and ease of using test questions for automated assignments with e-Learning tools, the question arises how to make the results most efficient. Supervising students in class during the test seems indispensable, but does not by itself ensure effective test outcome.

A popular belief between teachers is that test assessments measure students' knowledge and skills. Sometimes student general interest or involvement in the topics taught is also mentioned. By looking closer [1,5,6], more factors can be defined which can influence test results:

1. stress during test,
2. question wording, being misleading or unclear,
3. question relevance to the topic,
4. question matter coverage in training material provided,
5. time used for student preparation,
6. student's mental capabilities,
7. student internal motivation, which is influenced by his or her individual interests, and involvement in course,
8. luck in guessing answers.

The above can be summed up in somewhat non-constructive statement that the tests only measure students' proficiency of answering the test questions in the given test conditions.

To get results more closely related to student's performance, other factors should be ruled out.

Stress can be managed by e.g. good test organization, proper time setting, and ensuring smooth working of computer and network equipment during the test.

But items 2,3, and 4 from the list above remain valid. Al these items are related to the quality of questions themselves. It is common and very easy to blame students for lack of their preparation, when the results of the test are unsatisfactory. But it must be remembered that questions are the measurement equipment used during the test, and that their low quality may strongly and negatively influence the results.

A histogram in Figure 1 shows speculative results of a test containing multiple questions. It is assumed that

- questions have the same weight,
- the total score sums up to 5,
- each question has five answer options, of which just one is correct.

In other words the chances of guessing a single question is 1/5=0.2.

Assuming further a disastrous low quality of questions, every student would find himself or herself forced to guess the answer to every question. This would result in a Gaussian distribution of final test scores, with mean value at 1 and standard deviation of 1.

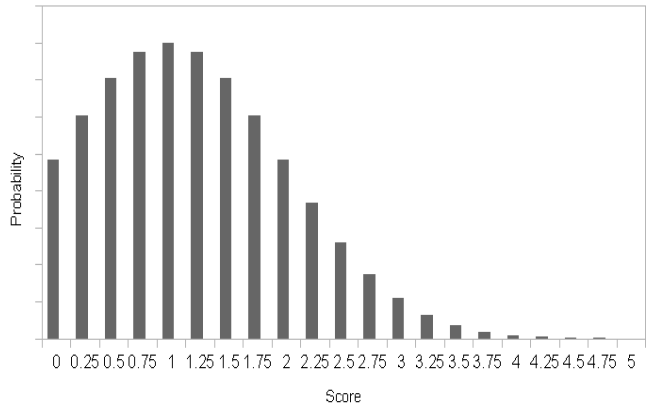

Fig. 1. Speculative results of a test in which students were guessing answers to all (100%) questions

The distribution of scores from Figure 1 may seem innocent at the first view, as it is widely accepted that student score distribution in tests should provide a Gaussian. But such distribution may suggest that most students failed to learn, despite its true origin from the assumptions specified above.

The second problem with the distribution from Figure 1 is that the bulk of students has scores assigned very close to each other, which makes it hard to distinguish individual student performance and to appoint a fair grade.

The third problem with the curve from Figure 1 is that most students will not be well-motivated to continue learning after getting grades resulting from such a distribution. This touches the topic of the goal of grading. One of the goals may be purely administrative, i.e. to assess if the student knows enough to pass. But grading during the semester can serve to motivate students. It is widely believed [1] that positive scoring increases internal motivation more effectively, than negative scoring increases the external motivation. Additionally, of these two, external motivation is considered inferior.

Therefore the curve will not be very usable from any of the presented tree points of view.

An easy shift of the curve is possible; assume 20% of the questions are obvious to everybody who participates in the test. Therefore everybody answers these 20% of questions correctly, and everybody receives a 20% increase in his or her score. This effectively shifts the distribution to the left by 1 (Fig.2).

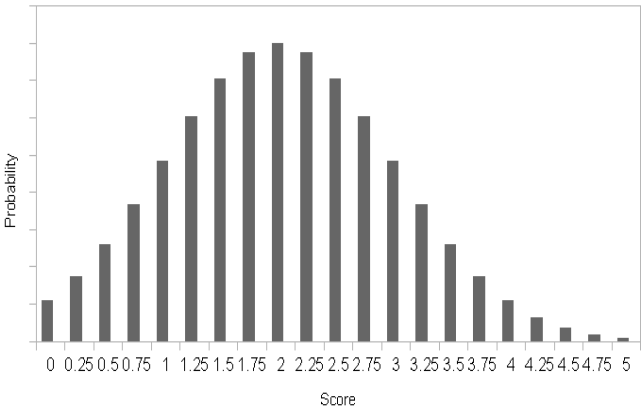

Fig. 2. Speculative results of a test in which students were guessing answers to 80% of question, while the remaining 20% of questions were obvious for everybody

Another belief which seems to be very common between teachers is that questions are mainly differentiated be their hardness, i.e. as being either easy, or middle, or hard.

Such base for differentiating questions leaves aside an important aspect: for whom are the questions easy or hard? The questions are rarely equally easy or hard for every student; notable exceptions being when nobody knows the answer or when everybody knows the answer.

From author's talks with peer teachers, histograms are done rarely, but score histograms as the ones presented in Figure 1 and Figure 2 are encountered and accepted in academic teaching. Mostly they are accompanied by comments about students failing to learn. But as presented in the above assumptions, both distributions have been obtained due to low question quality, and are not related to students' performance at all.

Histograms based on student test results are automatically generated in Moodle. With the increasing use of this tool for education, histograms will probably become more common. While histograms alone fall short of telling if the unsatisfactory score was caused by low quality questions or by weak student learning efforts, they may give rise to other interesting issues like what could be the reason for such apparent lack of correlation in teaching.

3. SOLUTION

Let's assume the following:

- Domain of knowledge and skills can be exhaustively covered by a set of N facts, to which N questions are connected in a one-to-one relation, so that the correct answer to a question represents the complete fact, accurately.
- The above set of facts is denoted D.
- Students possess knowledge of some facts, that is denoted K. This is the subject knowledge gained during the course or otherwise, and may not be entirely aligned with course content.

- There is a possibly empty common set C of domain knowledge D, covered by students' knowledge K:

$$C: D \cap K \quad (1)$$

3.1. Question coverage

Let's note that selecting to test the questions pertaining to facts in C results in a chance to some knowledgeable students to answer correctly by their mental capabilities. Selecting to test the questions out of C forces all students to guess.

Because score from all questions in a test are summed up together, the above described situation would lead to a distribution (Fig.3) flattened and shifted to the right, compared to the one in Figure 1. This is because some students will still get their score only by guessing, while others will get to different extent some additional score from the other questions that they know the answers for.

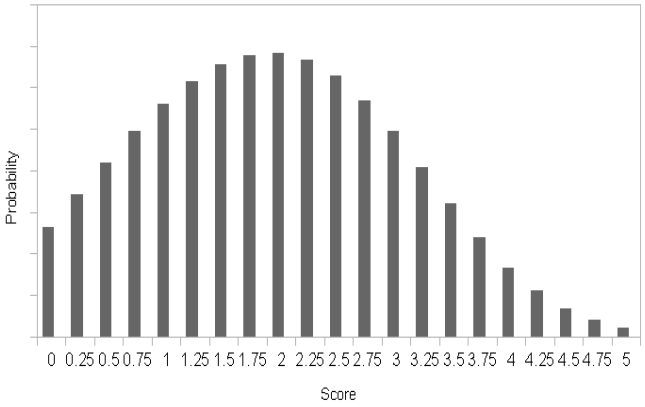

Fig. 3. Speculative results of a test in which students were guessing answers to part of the questions

It is out of the scope of this paper to deliberate why the set C cardinal number may be small relatively to the cardinality of D. From authors talks with peer teachers, this may be caused by not defining the scope of teaching clearly, e.g. by not providing teaching material covering the entire subject domain which is later verified in test. Of course not just the scope, but also the form of providing the teaching material is important. Just saying something during the lecture seems too volatile to call this form of knowledge transfer reliable. Knowledge transfer by speech is prone to missing, or misunderstanding, or not associating correctly with the rest of the facts – and therefore to quick forgetting

Another issue may be the lack of motivation in students to learn even when faced with a possibility to fail a test. This can have its own many explanations, not further researched here. Without defining closer what a sufficient level of motivation is or should be, it is assumed here that students in the population are generally not ill-motivated.

Therefore the basis for making good tests should be to clearly define and document D, and to select questions which only pertain to facts in D. When questions for the test cover only part of D, what is usual due to the sheer potential cardinality of D, students get additional difficulty getting right grades, as their knowledge form the set D not covered by questions is not credited to them. Therefore when only selected sub-set of domain D is enclosed in test questions, students knowing all the facts from D retain the advantage of fair grading, while the students not knowing all the facts from D get a grade which do not precisely reflect their capabilities. The grade will be higher if these weaker students happen to know the sub-set of D, or their grade may be lower if they happen to know fact from D which are out of this sub-set.

A solution to this is to make many smaller-scope tests, instead of a single big one. Besides of the obvious advantage of motivating students more often, and making their learning more effective in smaller chunks, this results in a more precise measure of individual student's performance.

3.2. Number of questions

It is usual that when there are N questions in the test, then correct answering of only S_i of them is sufficient to get a given i-th positive grade, $S_i<N$. This means that it is sufficient for a student to only know by heart answers to S_i-G questions, and to guess the remaining G answers, $G<S_i$, to get a given i-th grade. For G=1 this means to guess 1 out of the remaining $N-S_i$ questions. When probability of guessing a single question is 0.2, the probability P_G of getting any higher grade, not necessary only the (i+1)-th grade, with the help of guessing can be computed as:

$$P_G = 1 - P_{\text{not guessing}} = 1 - (1 - 0.2)^{(N-Si)} \quad (2)$$

It should be noted from (2) that getting a passing grade by a student who does not deserve this is most probable, as $N-S_1$ has the bigger value for the lowest grade, because S_1 is the lowest. Therefore it is quite easy for an unqualified student to pass. Taking an example test of 10 question with passing score of 60% i.e. requiring 6 correct responses to questions to pass, the above (2) means that there is a probability of over 0.67 to pass for a student who knows only 5 answers. Such a quite high probability value means that a majority of students who do not qualify due to not knowing one answer will pass anyway. It must be taken into account that not all the non-qualifying students will be just one answer away form passing, of course, but similar calculations can be conducted for G>1.

The above stresses the importance of using a sufficient amount of questions, corresponding to the grade thresholds, to get reasonable probability of grading students fair.

3.3. Question quality

Deciding whether a given score distribution is caused by students low level of knowledge D or by low questions quality, requires additional data. Moodle offers such data in the form of question statistics [7]. There are two parameters most useful in judging question quality:

- question ease, expressed as the rate of students getting this question correct
- question discrimination efficiency, which is based on the correlation between individual students' scores and the scores resulting from the single question under investigation [8].

4. CASE STUDY

Both the above question quality indexes must be considered in concert, to convey useful information. They inform for whom the questions were easy or hard: for the good students or for the weak ones. From author experience, this kind of statistics is very effective in identifying faulty questions. Creating high quality questions is much easier when these individual measurements are available, as they

help to find out which part of the quiz creation best practices [5,6] should be addressed.

From author experience, entry test questions are of weak quality if they concentrate on logical associations, deeper understanding, or ability to apply knowledge. Such questions should rather be oriented on checking the memorization of facts. By using Moodle metrics to identify faulty questions, author managed to drive the histograms to better serve the grading goals, e.g. to produce automated grading which motivates students on smaller assessments, like entry tests (Fig.4a), or to obtain. good student differentiation for final grading (Fig.4b).

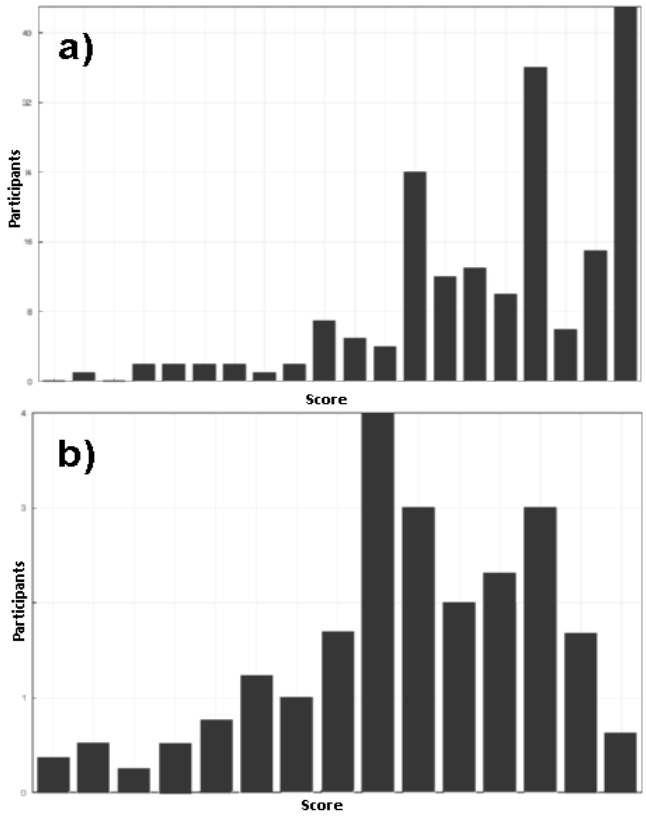

Fig. 4. Real life results of a) an entry test used to motivate students; b) a final test used to differentiate student achievements

5. CONCLUSION

Massive courses, besides the many challenges they introduce, provide also a lot of data which can be turned into useful information

One of the outcomes is that questions are not simple easy or hard, but that they usually exhibit different ease for different students..

The next observation based on the above is that questions have their quality, and that the answers to questions generally do not measure students knowledge, nor skills, nor involvement – without further efforts.

Question creation is a complex task; it is oversimplified to think that someone knowing well the subject matter will be able to create high-value questions only on this basis. Also the value of monitoring and refining question quality can not be overestimated.

Grading goals may be different – smaller assessment could be graded in a way which helps to motivate students, while the decisive assessment should probably provide different grading. Both goals are achievable with the right selection of questions.

6. BIBLIOGRAPHY

1. Gronlund, N. E., *Measurement and evaluation in teaching*, Prentice Hall College Div, 6th Ed., Chap.6-9, ISBN:9780023481116, 1990
2. Opaliński, A., MARKING LARGE AMOUNTS OF STUDENT ASSIGNMENTS, eTEE, e-Technologies in Engineering Education Conference, Gdansk, Poland, 2014
3. Lori Breslow, David E. Pritchard, Jennifer DeBoer, Glenda S. Stump, Andrew D. Ho and Daniel T. Seaton, *Studying Learning in the Worldwide Classroom: Research into edX's First MOOC*, Research & Practice in Assessment, Vol.5, 2013, available online, URL: http://www.rpajournal.com/dev/wp-content/uploads/2013/05/SF2.pdf, DOA:25.02.2014
4. Open Source Moodle Project home page, available online, URL: https://moodle.org/ , DOA: 25.02.2014
5. *How to Judge the Quality of an Objective Classroom Test*, Technical Bulletin #6, Evaluation and Examination Service, The University of Iowa, available online, URL:http://www.uiowa.edu/~examserv/resources_fees/Technical_Bulletins/Tech%20Bulletin%2006.pdf , DOA:25.02.2014
6. *Improving Your Test Questions*, Center for Innovation in Teaching & Learning, University of Illinois, available online, URL: http://cte.illinois.edu/testing/exam/test_ques3.html , DOA 25.02.2014
7. Moodle Documentation: *Quiz statistics report*, available online, URL: http://docs.moodle.org/24/en/Quiz_statistics_report , DOA: 25.02.2014
8. Moodle Documentation: *Quiz statistics calculations*, available online, URL: http://docs.moodle.org/dev/Quiz_statistics_calculations , DOA: 25.02.2014

WPŁYW ZESTAWU PYTAŃ NA WYNIKI UZYSKIWANE PRZEZ STUDENTÓW

Pojawienie się narzędzi e-Learningu umożliwiających automatyczne ocenianie wyników testów prawdopodobnie zwiększy wykorzystanie testów w edukacji technicznej. Te same narzędzia pozwalają mierzyć jakość pytań testowych. Tworząc odpowiednie zestawy pytań, można uzyskiwać odmienne cele. Artykuł zawiera teoretyczne i praktyczne przykłady histogramów punktacji testowej.

Słowa lluczowe: jakość pytań testowych, testy, e-Learning.