Influence of negative marking in online testing

Cristian Răzvan USCATU, Bucureşti, Romania, cristiu@ase.ro

Education reorientation towards the electronic, computerized environment naturally leads to transferring evaluation activities to the same environment. Beside the general trend of the society, use of Computer Aided Assessment is also required by other objective factors. Among them we have the huge raise in student numbers leading to the need to test and evaluate many students in a short time, in a coherent manner, while the numbers of the personnel administering the testing did not raise or raised much slower. Computers can perform this task removing any suspicion of subjective evaluation. Downside of this is that computers are machines that perform preprogrammed tasks. They lack the ability to understand free answers therefore testing has to be redesigned in a compatible manner. Also, some algorithms should be implemented to prevent and detect wrongful conduct from those undertaking tests.
Keywords: *computer aided assessment (CAA), online testing, negative marking, multiple-choice questions.*

Premise

Computer aided assessment has evolved out of the need to standardize tests and evaluate many students within a short period, as objectively as possible. Free answers, where evaluation is harder and many times subjective, are being replaced by standardized question where final answer may be expressed as "student has given the correct answer" or "student did not give the correct answer". This turns evaluation into a routine activity, suited for computers rather than humans. Once a testing system is implemented, it may be used extensively, not only for final examination, but also for partial examinations or formative testing. Formative testing improves the education process, identifying weak spots and correcting them on the way.

There are advantages and disadvantages when using CAA ([Stephens, 1997]). Beside allowing a coherent and standardized method of testing large numbers of students in a short period, CAA may be used repeatedly, on demand, to guide the learning process, highlighting strong spots and weak spots. It allows automated quick monitoring of large numbers of students, which frees the professors' time for other, more important formative activities like direct interaction with students. Computer systems allows using large databases of tests; ad-hoc combinations of questions and automated reordering of questions and answers helps prevent cheating on exams. This tests cover more of the studied subjects than a classic test. Internet and intranets allow simultaneous testing of many students and platform independence. Also, CAA reduces costs and resource consumption (mainly paper and time).

Disadvantages derive form the computer limitations. Since computers can only perform preprogrammed tasks and cannot understand free answers, tests must be redesigned. They rely on memorization and recognition. Not all subjects are suitable for this kind of testing, some have to continue using classic evaluation. Using the internet brings the problem of data security and lack of control on what the student is really doing during the test. Tests and questions must be carefully created to avoid including clues that point to the right answer. Creation of "good" questions may be difficult and time consuming. The possibility of pure random guessing may lead to the use of negative marking schemes when grading the tests.

CAA may use a large variety of question types ([CIAD, 2003b] and [Hallam, 2003] present a comprehensive list), but the most widely used is the multiple choice question (MCQ). This kind of questions present the several answers to the student (usually 4 or 5) of which only 1 is correct. A person may randomly guess the answers to all questions and still get some of them right, when a large number of questions are used. Questions with

answers randomly guessed add undeserved points giving a false image on the student. This is known as the "monkey score" ([Leicester, 2002] - a monkey may choose answers and get the same score).

There are strategies to discourage this kind of cheating, usually known as negative marking schemes. They involve deducting points for each wrong answer or adjusting the final grade. Some fields are suitable for this kind of marking, strongly discouraging blind guessing, while others encourage using partial knowledge in attempts to discover the right answer.

Negative marking schemes

Negative markings may be applied either on each question or on the entire test ([Leicester, 2002], [Pettigrew, 2001], [Freewood, 2001]).

On question level, without negative marking, a correct answer brings 1 point while a wrong answer brings 0 points (no gain, no loss). Sometimes pure random guessing may bring enough points to pass an exam. For example, on a test with only 2 choices for each question, on average random guessing will produce 50% correct answer, usually enough to pass. In other cases, random guessing may add enough points to those deserved to pass the exam. The goal of negative marking is to eliminate the undeserved points so that a perfectly prepared student will get 100% points while a student that randomly guesses all answers will get 0%. Other student will achieve some percents, according to the level of their knowledge.

For a test with only 2 choices for each question, a point may be awarded for a correct answer and a point deducted for each wrong answer. On pure random guessing, 50% correct answer will be cancelled by the 50% wrong answers and final score will be zero. If there are more choices, this marking scheme does not work right. Suppose there are 5 choices for each question, random guessing will produce 20% correct answers and 80% wrong answers. Using previous scheme, this leads to an average -60% score for a student that randomly guesses all answers. In order to comply with the above stated goal, correct answers should bring 4 points, not 1. In a similar manner, a test with 4 choices should award 3 points for a correct answer, a test with 3 choices should bring 2 points for a correct answer.

If n is the number of choices (the same for all questions, of which 1 choice is correct and $n-1$ are wrong) and C is the number of points for a correct answer, the correction for each wrong answer selected is $I = \frac{C}{n-1}$. If no answer is selected, the gain is null (zero points). Tests may explicitly include this choice ("I do not know").

Another way to use negative points is to associate a certainty to each answer (for example, ranging from 1 to 5 – where 5 is the number of choices). If the selected choice is correct, the student gains a number of points equal to the certainty chosen for the answer, otherwise he is deducted a number of points equal to the chosen certainty. Deduction may be more severe (even double) in some domains (for example, a wrong diagnostic with a high certainty in medicine is heavily discouraged).

Another scheme associates uses an order of preferences. Students must sort the choices in the order of preference. Each answer has a number of points associated, depending on its position in the list (for example, if there are 5 choices, first choice is worth 4 points, second choice is worth 3 points and so on; last choice is worth 0 points). The number of points gained depends on the position of the correct answer in the list. This method is largely used in online testing, where the student may try again until he/she chooses the right answer.

Liberal tests seek to encourage students to apply partial knowledge. They allow the student to choose more than one answer (although only 1 answer is correct). This way the student may eliminate some answer based on partial knowledge and get some partial points instead of random guessing which may bring either zero or 100% points. For a question with 5 choices:

- if the correct answer is selected (and only this one) the student gains the full points for dismissing 4 out of 4 wring answers.

- if two answers are chosen, including the correct one, 3 out of 4 wrong answers were dismissed which means 75% points.
- selecting 3 answers (including the correct one) means 2 out of 4 wrong choices dismissed, so 50% points are awarded.
- 4 answers selected means 25% points.

These percentages are used if all answers have equal preferences.

When no negative marking schemes are used on grading each question, final score may be adjusted. If N is the unadjusted score and n the number of choices, the adjusted score is $A = \frac{n * N - 10}{n - 1}$. The effect of this adjustment is an increase in the minimum passing score. For a test with 5 choices, a student must achieve a grade of 6 (unadjusted) in order to pass the exam with a grade of 5 (adjusted). If there are 4 choices, he must gain a grade of 6.25 in order to pass. Common grades are adjusted as follows (test with 5 choices):

unadjusted	1	2	3	4	5	6	7	8	9	10
adjusted	-1.25	0	1.25	2.5	3.75	5	6.25	7.5	8.75	10

Perfectly prepared students are not affected by this adjustment. Also, the better they are prepared, the less they are affected by the adjustment. This method does not change the order of the students' grades but enlarges the gaps between them. Students with average luck would score 20%, which is adjusted to zero. Students that are not so lucky may score less than average when guessing which leads to negative grades after adjustment. They may be replaced with the minimum grade (usually 1).

When negative marking are being used, negative scores may result and scores are spread over a larger range. These scores must be brought within the normal range (1-10).

For example, a test with 100 question with 5 choices, top possible score is 400 and lowest possible score is -100. Let's suppose a student correctly answers 60 questions, random guesses the answers for 20 question and for the rest of them eliminates 3 wrong answer using partial knowledge, guessing the correct answer of the 2 remaining choices. For the first 60 questions, he will gain 4*60-240 points. Of the 20 question where random guessing was used, in average 20% (4 questions) will be correct, the rest will be wrong. The corresponding score will be 4*4-16*1=0. For the last 20 questions, in average half will have the right answer and the other half will be wrong. The score will be 10*4-10*1=30 points. Total score is 240+0+30=270.

Normalization may consider the whole possible range of score or only the top possible score. For the 270 points, the grade will be:

- $\frac{270-(-100)}{400-(-100)} * 10 = 7.4$, if the entire score range is used;
- $\frac{270}{400} * 10 = 6.75$, if only the top possible score is used.

In the second case, students with less luck when guessing will score less than average, which means negative scores. Those cannot be turned into grades. One way is to simply replace them with the minimum grade (1).

Influence on test results

Negative marking schemes were tested on 236 students undertaking multiple-choice tests consisting of 40 questions each. No negative marking scheme was really applied, but the actual results have been used for this research.

Tests consist of 40 multiple choice questions, with 5 choices each. Each correct answer is worth 2.25 points, a total of 90 points available. 10 points are awarded to everyone, thus final grades ranging from 1 to 10 (10 to 100 points, normalized by division to 10).

With no negative marking schemes, 224 students (94.92%) passed the exam, 12 failed. Grades ranged from 3.48 to 9.33, with an average grade of 6.85. This is designated as set 1.

When using negative marks, a correct answer is awarded 4 units (4*2.25=9 points) and for wrong answer 1 unit (2.25 points) are deducted. Scores are brought within the normal range first by reporting to the full range, than by reporting only to the top score.

Set 2 refers to the grades computed using the whole range of scores as reference. Grades in this set are identical to the ones in set 1.
Set 3 refers to the grades computed using only the top score as reference. In this set grades are lower than set 1. They range from 1.84 to 9.16 with an average grade of 6.06. In this set only 182 students pass the exam, while 54 fail.
The last set (number 4) adjust grades using this formula, described above: $A = \frac{n * N - 10}{n - 1}$. Final grades are identical to those in set 3. Figure 1 shows the grades for the two pairs of sets. Sets 1 and 2 are represented by the top line.
The most visible impact of using the adjustment on the grades is lowering the final grades. This leads to more students failing the exam. For the top students, differences are insignificant: 0.17 points for the top grade (9.33 versus 9.16). The difference increases as the grades become lower reaching 1 point for the first failed student and 1.63 points for the lowest grade.
This means top students are not really affected, while those with poor training will find it more difficult slipping past the exam. This may be an incentive to make everybody study harder and rely less on guessing.
Of course, knowing that a pass grade is harder to get could also lead to resistance against this method of grading tests. Fear of negative markings could make some students to perform worse than they are really trained.

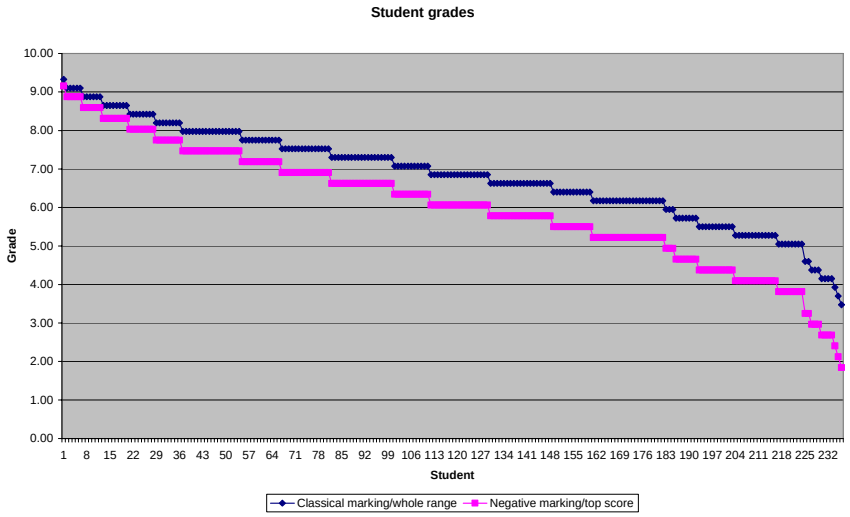

Fig.1. Adjusted versus unadjusted grades

References

[1] [CIAD, 2003a] Centre for Interactive Assessment Development, Don Mackenzie, Dave O'Hare, Chris O'Reilly, Helen Wilkins – *Assessment marking strategies*, www.derby.ac.uk/ciad/dev/logical.htm
[2] [CIAD, 2003b] Centre for Interactive Assessment Development – *Question Styles*, www.derby.ac.uk/ciad/Questions.htm
[3] [Hallam, 2003] Sheffield Hallam University, Learning and Teaching Institute – *Computer Assisted Assessment – practical guide*, 2003 www.shu.ac.uk/services/lc/cmeweb/grant/ltiweb/08/caapracguide.html
[4] [Freewood, 2001] Madeleine Freewood – *Negative Marking*, 2001, www.shu.ac.uk/services/lti/resources/caapracticalguidance/negativemarking.html
[5] [Leicester, 2002] University of Leicester – *The Castle Toolkit – Designing and managing MCQ's, chapter 4: Score and Statistics*, 2002, www.le.ac.uk./castle/resources/mcqman/mcqcont.html, www.uct.ac.za/projects/cbe/mcqman/mcqchp4.html
[6] [Pettigrew, 2001] Mark Pettigrew – *Random guessing on multiple choice questions*, www.shu.ac.uk/services/lti/people/mp/mcq
[7] [Stephens, 1997] Derek Stephens, Janine Macia – *Results of a Survey into the use of Computer-Assisted Assessment in Institutions of Higher Education in the UK, January 1997*, www.lboro.ac.uk/service/ltd/flicaa/downloads/survey.php