Initial correction versus negative marking in multiple choice examinations

Tanja Van Hecke
Ghent University
Faculty of Engineering and Architecture
Ghent, Belgium
Tanja.VanHecke@ugent.be

Abstract

Optimal assessment tools should measure in a limited time the knowledge of students in a correct and unbiased way. A method for automating the scoring, is multiple choice scoring. This paper compares scoring methods from a probabilistic point of view by modeling the probability to pass: the number right scoring, the initial correction and the negative marking method. We will compare the probabilities for students to pass when their assessment is translated into a score by means of the negative marking and the initial correction method. Moreover, given a knowledge level of the student, the variance of this probability will be discussed for both methods.
Keywords: multiple choice; test fairness; scoring

1 Introduction

Multiple choice testing becomes more popular at Belgian universities due to the increase of the number of student enrolments and the need for multiple periodical assessments. As MC testing fits in computer aided assessment techniques, the time and effort invested in grading can be kept down [8]. A considerable drawback is the difficulty to separate the better and the moderate students, so a correcting scoring method is required. The number right scoring method (NR) [5], where the total score is the sum of positive scores for the correct answers, can be corrected into the negative marking method (NM) where wrong answers are penalized [5]. Holt [4] already described how for a negative marking scheme the benefits/detriments of guessing depend upon the severity of the penalty of an incorrect answer relative to the level of reward for a correct answer and the number of options from which the students have to choose. In literature many authors agree that NM used to determine the scores for a multiple choice (MC) examination, unfairly disadvantages the students that are reserved towards guessing. Discouraged by the possible penalty, they may choose for the secure option of leaving the question blank. This problem is not associated with NR, a method where no penalty exists but where bias is a consequence of the arbitrary guessing. A better alternative method is the Initial Correction method (IC), a method that provides an initial fixed correction which is augmented by each correct answer, which we will analyse in this paper and is related to methods described in [7]. Belal and Ammar [3] suggested the introduction of certainty/confidence levels for decision questions of the true/false type. They ask the students to rank the test items relative to each other according to his/her confidence level in answering the item correctly. The test is graded according to the selected rank sequence. There is no penalty for wrong answers. A step-size is used to determine changes in the students confidence level based on the number of incorrect answers. A reduction function is used to determine the scores for correct answers at the different confidence levels. A disadvantage of the method of Belal and Ammar is its complexity, which makes it difficult to communicate the way of scoring to the students. NM as well as IC have the advantage that you can explain it easily to students. McGinty [6] encourages the inclusion of a variety of qualitative methods such as naturalistic observation and think-aloud studies, as well as quantitative methods that draw on existing theory and research outside the field of measurement.

The comparison in this paper is based on a probability study that will allow us to make certain claims about the optimal scoring of multiple choice tests. To analyze scoring methods we introduce parameters as in Table 1. We will include q in our analysis, i.e. the number of correct answers without guessing, which reflects students' knowledge. The value of q as well as the number of wrong answers due to students' wrong understanding of the course, are difficult to validate in practice. The latter is not included in our analysis as we focus on the danger of passing thanks to correct guesses.

Table 1: Parameters of a multiple choice examination.

parameter	definition
n	number of questions
a	number of alternatives for each question
c	number of correct answers
w	number of wrong answers
q	number of correct answers due to knowledge (not to guessing)

2 Scoring methods

2.1 Number right

No correction for guessing is included with the number right scoring method as the total score is simply

$$s_{corr} = c. \tag{1}$$

To visualize the probability distribution of the scores in case of a MC examination, we consider an examination with $n = 20$ and $a = 4$. Figure 1 shows the distribution of the score s ($s \in \{0, 1, 2, \ldots, 20\}$) when a student does not know the answer for any of the questions and makes a guess for all of them. That kind of student will pass ($s \geq 10$ at Belgian universities) with probability 1.39%, using the formula

$$\sum_{j=n/2}^{n} C_j^n \left(\frac{1}{a}\right)^j \left(1 - \frac{1}{a}\right)^{n-j}. \tag{2}$$

When the student has partial knowledge as he knows the correct answers to some of the questions but not to all of the questions, a shifted plot with less variance is generated as in Figure 2 for $q = 5$ (left) and $q = 10$ (right). The variability is influenced by three factors: n, q and a (see Figure 2 and 3).

2.2 Negative marking

A penalty for guessing is provided in NM as the total score s_{corr} for the examination is contributed as

$$s_{corr} = c - w\,\frac{1}{a-1}. \tag{3}$$

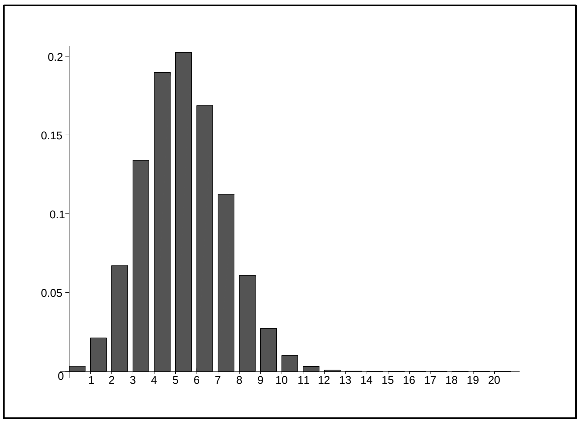

Figure 1: PDF of score with NR ($n = 20$, $a = 4$, $q = 0$).

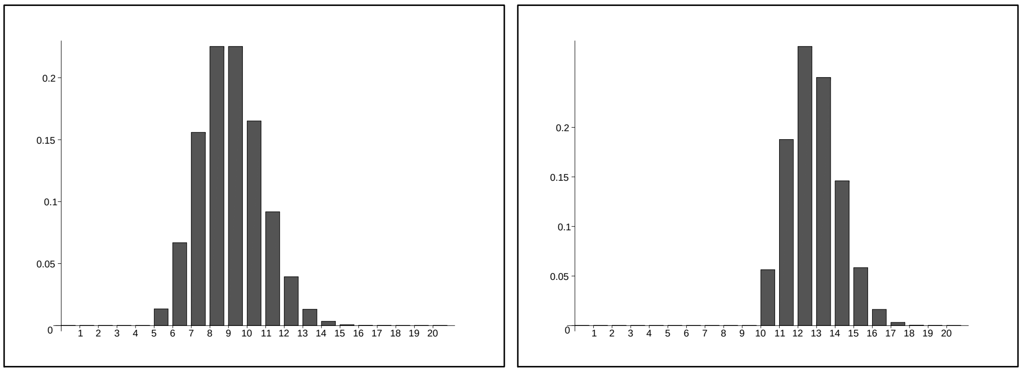

Figure 2: PDF of score with NR ($n = 20$, $a = 4$) for $q = 5$ (left) and $q = 10$ (right).

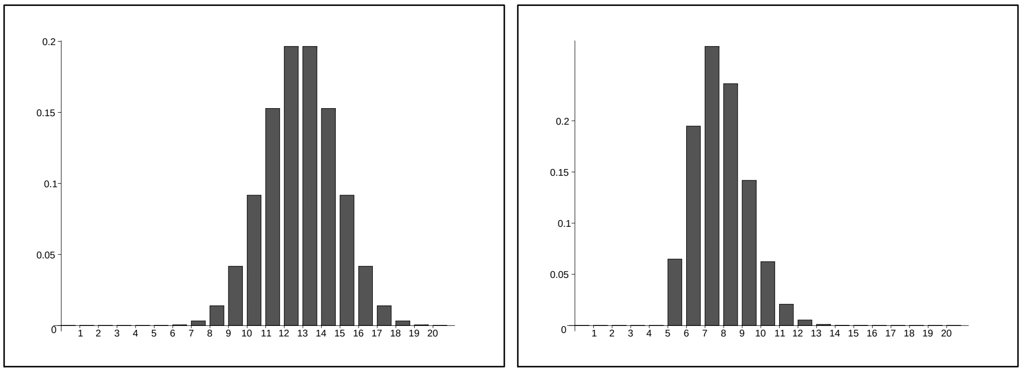

Figure 3: PDF of score with NR ($n = 20$, $q = 5$) for $a = 2$ (left) and $a = 6$ (right).

When analysing NM, it is assumed that knowledge is binary: students either select the right answer or pick one randomly. However this is not always realistic: many students do not know the correct answer to a question but can eliminate some of the distractors as being incorrect. [2]. When the binary assumption holds, the expected value of a pure guess is equal to the expected value of leaving the answer blank. When the binary assumption is not valid and a penalty of $1/(a-1)$ is used, the expected value of guessing is larger than the value of leaving the answer blank. Due to the guessing penalty the scores are biased against risk-averse students [1].

2.3 Initial correction

We propose a generalisation of the total score described in [7] as

$$s_{corr} = -\frac{n}{x} + c\,(1 + \frac{1}{x}), \tag{4}$$

which we will evaluate for different values of the parameter x. The starting point is a negative value $-\frac{n}{x}$. Each correct answer is worth $1 + \frac{1}{x}$ points to reach a maximal score of n when all questions are answered correctly. As no penalty for guessing is provided, the students get an incentive to guess, which increases the measurement error. Remark dat (4) equals (3) if $x = a - 1$ in combination with no blanks ($w = n - c$) for NM. Moreover the expected value of the score for someone who randomly guesses the answers for all the questions is 0 if $x = a - 1$ as

$$\begin{aligned} E[s_{corr}/(\mathrm{IC},\, q=0)] &= -\frac{n}{a-1} + c_{\text{when } q=0}\left(1 + \frac{1}{(a-1)}\right) \\ &= -\frac{n}{a-1} + \frac{n}{a}\,\frac{a}{(a-1)} \\ &= 0 \end{aligned}$$

3 Probability to pass

As guessing is not punished with the IC method, all questions will be answered. The enlarged number of answered questions will create a larger variability in scores with IC compared to NM, given a knowledge level of a student. The questions that the student would have left blank with NM, are answered by guessing when IC is used. The increased variability has the disadvantage that a student able to answer correctly half of the questions, can fail in a system of IC, where he would have passed in a system of NM.

Students' major aim is passing. The probability to pass with IC is given by (5), where b (b_{NM} for NM and b_{IC} for IC) denotes the number of correct guesses among the $n - q$ questions wherefore the answer is not known by the student.

$$p_{pass} = 1 - \sum_{j=0}^{b-1} C_{n-q}^{j} \left(\frac{1}{a}\right)^{j} \left(1 - \frac{1}{a}\right)^{n-q-j} \tag{5}$$

As $n/2$ is the threshold to pass (at Belgian universities), the score built up with the scoring methods should reach $n/2$ at least:

$$\begin{aligned} \text{NM:} \qquad & \frac{n}{2} \leq (q+b) - (n-q-b)\,\frac{1}{a-1} \\ \Leftrightarrow \quad & b_{NM} = \left\lceil \frac{n\,a + n}{2\,a} - q \right\rceil \end{aligned} \tag{6}$$

$$\begin{aligned} \text{IC:} \qquad & \frac{n}{2} \leq -\frac{n}{x} + (q+b)\,(1 + \frac{1}{x}) \\ \Leftrightarrow \quad & b_{IC} = \left\lceil \frac{n\,x/2 + n}{x+1} - q \right\rceil \end{aligned} \tag{7}$$

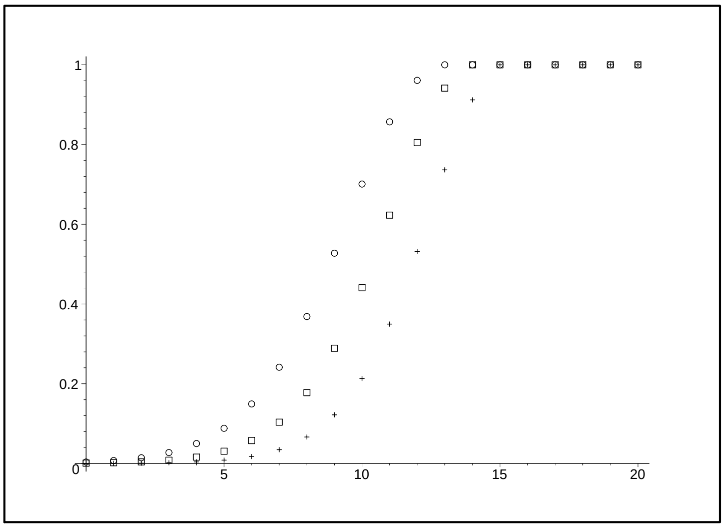

Figure 4: Probability to pass as a function of q for NM, IC$^{[a-1]}$ (box), IC$^{[a-2]}$ (cross) and IC$^{[a]}$ (circle) for $a = 3$.

or

$$b_{IC}^{[a-2]} = \left\lceil \frac{n\,a}{2\,(a-1)} - q \right\rceil \quad \text{for } x = a-2, \tag{8}$$

$$b_{IC}^{[a-1]} = \left\lceil \frac{n\,a+n}{2\,a} - q \right\rceil \quad \text{for } x = a-1, \tag{9}$$

$$b_{IC}^{[a]} = \left\lceil \frac{n\,a/2+n}{a+1} - q \right\rceil \qquad \text{for } x = a. \tag{10}$$

Remark that $b_{IC}^{[a-1]} = b_{NM}$.

Figure 4 ($a = 3$) and Figure 5 ($a = 4$) compare the probabilities for a student to pass as a function of the number of questions q wherefore the student knows the answer. It is assumed that the student makes a guess for all other questions. Obviously the less the student has to guess, the higher his probability to pass, but the probabilities to pass with IC are higher when the initial score $-\frac{n}{x}$ is higher, i.e. the value of x is larger. These figures and Table 2 also show that the variability increases when the number of alternatives a increases and that NM excels in reducing the variance. Remark that the sharp step function which changes from probability 0 to 1 at $q = 10$, representing the binary pass/fail system, is best resembled by the negative marking method and for higher values of a.

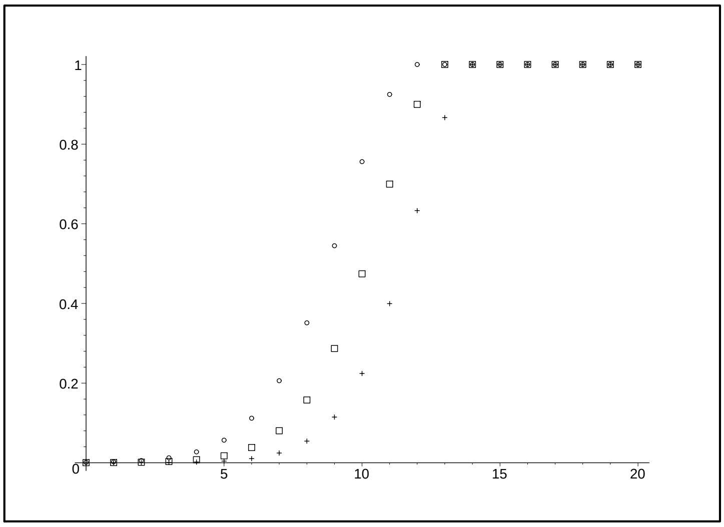

Figure 5: Probability to pass as a function of q for NM, $\mathrm{IC}^{[a-1]}$ (box), $\mathrm{IC}^{[a-2]}$ (cross) and $\mathrm{IC}^{[a]}$ (circle) for $a = 4$.

Table 2: Variance of the probabilities to pass with different scoring methods and number of alternatives

a	scoring method	variance of p_{pass}
3	NM	0.1901
4	NM	0.2025
3	$\mathrm{IC}^{[a]}$	0.1995
4	$\mathrm{IC}^{[a]}$	0.2115
3	$\mathrm{IC}^{[a-2]}$	0.1999
4	$\mathrm{IC}^{[a-2]}$	0.2111

4 Conclusion

The comparative probabilistic analysis of the scoring methods for multiple choice questions reveals that students have higher probabilities to pass when their assessment is translated into a score with the initial correction method with an initial correction higher than $-\frac{n}{a-1}$ compared to the negative marking method. The negative marking method excels by its lower variance for the probabilities to pass given a knowledge level of the student. A positive issue about both types of methods is their correction for the biased character created by guessing, which is not available with the number right scoring method.

References

[1] AVILA, C. & TORRUBIA, R. (2004) Personality, expectations, and response strategies in multiplechoice question examinations in university students: A test of Grays hypothesis. *European Journal of Personality*, **18**, 45–59.

[2] BLISS, L.B. (1980) A test of Lords assumption regarding examinee guessing behavior on multiplechoice tests using elementary school students. *Journal of Educational Measurement*, **17**, 147–153.

[3] BELAL, A.A. & AMMAR, D.F. (2010) Confidence-Based Assessment of Two-Alternative Format Tests. *Paper presented at the 5th Conference of Learning International Networks Consortium (LINC) MIT, Massachusetts, USA.*

[4] HOLT, A. (2006) An Analysis of Negative Marking in Multiple-Choice Assessment. *Paper presented at the 19th Annual Conference of the National Advisory Committee on Computing Qualifications (NACCQ 2006), Wellington, New Zealand.*

[5] KURZ, T.B. (1999) A review of scoring algorithms for multiple-choice tests. *Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.*

[6] MCGINTY, D. (2005) Illuminating the Black Box of Standard Setting: An Exploratory Qualitative Study. *Applied Measurement in Education*, **18(3)**, 269–287.

[7] PRIHODA, T.J., PINCKARD, R.N., MCMAHAN, C.A. & JONES, A.C. (2006) Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. *Journal of Dental Education*, **70(4)**, 378-386.

[8] WEISS, B., GRIDLING, TRODHANDL, C. & ELMENREICH, W. (2006) *Embedded Systems Exams with True/False questions: A case study (Report No. 5).* Institut fur Technische Informatik, Wein.

5 Biographical note

Tanja Van Hecke obtained a master degree in applied mathematics in 1995 at the Ghent University. In 1998 she finished her doctoral thesis in numerical analysis in the field of solving ordinary differential equations. For several years she worked at the faculty of applied engineering sciences at the University College Ghent where she was part of the department of mathematics and statistics. In 2013 this faculty became part of the faculty of Engineering and Architecture at the Ghent University.

Contact:
Tanja Van Hecke
Ghent University
Faculty of Engineering and Architecture,
Voskenslaan 270
B-9000 Ghent, Belgium
Email: Tanja.VanHecke@ugent.be