Balanced Scoring Method for Multiple-mark Questions

Darya Tarasowa and Sören Auer
Universität Leipzig, Postfach 100920, 04009 Leipzig, Germany

Keywords: Multiple-choice Questions, Multiple-mark Questions, Scoring Methods.

Abstract: Advantages and disadvantages of a learning assessment based on multiple-choice questions (MCQs) are a long and widely discussed issue in the scientific community. However, in practice this type of questions is very popular due to the possibility of automatic evaluation and scoring. Consequently, an important research question is to exploiting the strengths and mitigate the weaknesses of MCQs. In this work we discuss one particularly important issue of MCQs, namely methods for scoring results in the case, when the MCQ has several correct alternatives (multiple-mark questions, MMQs). We propose a general approach and mathematical model to score MMQs, that aims at recognizing guessing while at the same time resulting in a balanced score. In our approach conventional MCQs are viewed as a particular case of multiple-mark questions, thus, the formulas can be applied to tests mixing MCQs and MMQs. The rational of our approach is that scoring should be based on the guessing level of the question. Our approach can be added as an option, or even as a replacement for manual penalization. We show that our scoring method outperforms existing methods and demonstrate that with synthetic and real experiments.

1 INTRODUCTION

Advantages and disadvantages of a learning assessment based on multiple-choice questions (MCQs) are a long and widely discussed issue in the scientific community. However, in practice this type of questions is very popular due to the possibility of automatic evaluation and scoring (Farthing et al., 1998). Consequently, an important research question is to exploit the strengths and mitigate the weaknesses of MCQs. Some systems (e.g. Moodle [1]) allow teachers to create MCQs with multiple correct options. This type of questions we will call multiple-mark questions (MMQs), to distinguish them from the conventional MCQs, where there is always only one correct option. Multiple-mark questions were already recommended by Cronbach (Cronbach, 1941). Other research (Ripkey and Case, 1996; Pomplun and Omar, 1997; Hohensinn and Kubinger, 2011) considers MMQs to be more reliable, when compare them with conventional MCQs. However, even though the advantages of MMQs are meanwhile widely accepted, up to our knowledge there are no balanced methods for scoring multiple-mark questions available to date.

One possible approach to score the MMQs is to use dichotomous scoring system. The dichotomous scoring awards the constant amount of points, when the question is answered correctly and zero points in a case of *any* mistake. However, the partial scoring is preferable to the dichotomous, especially in case of MMQs. (Ripkey and Case, 1996; Jiao et al., 2012; Bauer et al., 2011; Ben-Simon et al., 1997)

The second possible approach is to use the methods, developed for scoring the multiple true-false questions (MTFs). However, despite the possibility to convert the MMQs into MTFs, the studies (Cronbach, 1941; Dressel and Schmid, 1953) discuss the differences between two formats and show disadvantages of MTF questions compared to MMQs. In the paper we show that the differences prevent the applying methods developed for the MTFs to the MMQs scoring.

Another possible approach is to use the penalties, similarly to the paper-based assessment where the teacher can analyze the student answers and decide how much points she deserves. The method was proposed by Serlin (Serlin and Kaiser, 1978). For example, in Moodle a teacher has to determine what penalty applies for choosing each distractor. However, this work is an additional, unpopular burden for teachers, since not required in paper-based tests. Instead of asking the teacher, some systems calculate the penalties automatically. However, computer-based assessment opens additional possibilities to guess, for example choosing all options. Often the scoring algorithms do

[1] https://moodle.org/

Tarasowa D. and Auer S..
Balanced Scoring Method for Multiple-mark Questions.
DOI: 10.5220/0004384304110416
In *Proceedings of the 5th International Conference on Computer Supported Education* (CSEDU-2013), pages 411-416
ISBN: 978-989-8565-53-2

not take into account such ways of guessing.

Consequently, we are facing the challenge to find a scoring method, that is able to recognize and properly penalize guessing. Previously proposed algorithms suffer from imbalance and skewness as we show in 2.

The task to find the scoring method can be divided into two steps: to find a method to determine points for the correctly marked options(1) and to find a method to determine the penalty for the incorrectly marked options(2). For the first part a reasonable approach was proposed by Ripkey (Ripkey and Case, 1996). Thus our research aims to provide a method for the second part (determining penalties). We propose a general approach and a mathematical model, that takes into account the most common ways of guessing and behaves balanced at the same time.

Our concept is based on the assumption, that scoring can be based on the *guessing level* of the question. By guessing level we mean here (in partial scoring) the probability to obtain more than zero points. Each question is associated with a difficulty to guess a (partially) correct answer. To accommodate the difficulty level of guessing in the scoring method, we propose to determine the penalty only when a student marks more options, than the actual number of correct ones. We argue that our approach can be added as an option, or even as a replacement of manual designation of penalties. We claim that our algorithm behaves better, than existing ones and prove that with both synthetic and real experiments.

The paper is structured as follows: first, we discuss existing algorithms for scoring MMQs, then we describe our approach on conceptual and mathematical levels and finally we show and discuss the results of synthetic and real-life experiments.

2 RELATED WORK

There are several existing platforms, that use multiple-mark type of questions as well as several approaches to score them. We collected such approaches to describe, discuss and compare them. Existing approaches for scoring the multiple-mark questions implement four base concepts. In the section we describe the basic ideas, advantages and disadvantages of these concepts.

2.1 Dichotomous Scoring

This method is often used in paper-based quizzes, where the good quality of quizzes allows teacher to be more strict when score the results. As the aim of e-based quizzes is not only to score the results, but to catch the gaps of knowledge, the scoring of partially correct responses shows the actual knowledge of the student better. Also, dichotomous scoring does not show the accurate progress of the student. However, when dealing with multiple-mark questions dichotomous scoring almost excludes the possibility of guessing, that is why we use it as a standard of reference when evaluating our approach with real users.

2.2 Morgan Algorithm

One of the historically first methods for scoring the MMQs was described in the 1979 by Morgan (Morgan, 1979). For our experiments we use the improved algorithm, in accordance to which the scores are determined by the following algorithm:

1. for each option chosen which the setter also considers correct, the student scores $+(p_{max}/n)$, where n is a number of correct options
2. for each option chosen which the setter considers to be incorrect, the student scores $-(p_{max}/k)$, where k is a number of distractors.
3. for each option not chosen no score, positive or negative, is recorded regardless of whether the setter considers the response to be correct or incorrect.

However, the experiments show a large dependence between number of options (correct and incorrect) and amount of penalty, that indicates the skewness of the method (see 4.1).

2.3 MTF Scoring

Multiple-mark questions can be scored with the approaches developed for multiple true-false items. Tsai (Tsai and Suen, 1993) evaluated six different implementations of the approach. Later his findings were confirmed by Itten (Itten and Krebs, 1997). Although both researches found partial crediting to be superior to dichotomous scoring in a case of MTFs, they do not consider any of the algorithms to be preferable. This fact allows us to use the most base of them for our experiments.

MTF scoring algorithms imply that any item has n options and a fully correct response is awarded with full amount of points p_{max}. If the user did not mark a correct option or marked a distractor, she is deducted with the penalty $s = p_{max}/n$ points. Thus a student receives points for not-choosing a distractor as well as for choosing a correct option. This point does not fit perfect to multiple-mark questions because of the differences between two types (Pomplun and Omar, 1997; Cronbach, 1941; Frisbie, 1992). Our experiments (see 4.1) confirm the studies and show the skewness of the concept when deal with MMQs.

2.4 Ripkey Algorithm

Ripkey (Ripkey and Case, 1996) suggested a simple partial crediting algorithm, that we named by the author. In the approach a fraction of one point depending on the total number of correct options is awarded for each correct option identified. The approach assumes no point deduction for wrong choices, but items with more options chosen than allowed are awarded zero points. The Ripkey's research showed promising results in a real-life evaluation. However, later researches (e.g. Bauer (Bauer et al., 2011)) notice the limitations of the Ripkey's study. The main issue in the Ripkey algorithm is the not well-balanced penalty. We aim to improve the Ripkey's algorithm by adding the mathematical approach for evaluating the size of penalty.

3 BALANCED SCORING METHOD FOR MMQs

3.1 Concepts

As shown above, existing approaches do not solve the problem of scoring MMQs perfectly. Our concept is based on the assumption, that scoring can be based on the guessing level of the question. Thus, when a student marks all possible options, she increases the guessing level up to 1. In this case the student should obtain either the full amount of points (if all the options are considered to be correct by the teacher), or zero, if the question has at least one distractor. However, if a student did not mark any option, the score should be always zero, as we assume that all the questions have at least one correct option. Thus, the task is to find the correctness percentage of the response and decrease it with a penalty, if the guessing level was artificially increased by marking too many options.

Questions have the native level of guessing, and we propose to deduct the penalty only if after the student's response the guessing level increases. In other words, we determine the penalty only when a student marks more options, than the number of correct ones.

3.2 Mathematical Model

In this section we present the mathematical model, that can be used for its implementation.

3.2.1 Scoring the Basic Points

To score the basic points we use the approach, described by Ripkey. Below we present it mathematically in accordance with the following designations:

- $d \in \mathbb{R}, d \in (1..d_{max}]$ – difficulty of the current question, for our experiments we set $d_{max} = 5$
- $C \subseteq A$ – set of the *correct* options c_i for the current question, where A – set of the options a_j for the current question,
- $c_{max} = |C|, c_{max} \in \mathbb{N}$ – number of *correct* options for the current question
- C_{ch} – set of the *correctly checked* options
- $c_{ch} = |C_{ch}|, c_{ch} \in \mathbb{N}, c_{ch} \in [0, c_{max}]$ – number of *correctly checked* options for the current question
- $p_{max} = f(d) = d * K_{points}$ – maximal possible points for the current question, in our system we set $K_{points} = 1$
- p_c – points for the correctly checked option c. As we assume all the correct options have the equal weight, $\forall c \in C_{ch} | p_c = \frac{p_{max}}{c_{max}}$
- $p \in \mathbb{R} \wedge p \in [0, p_{max}]$ – the basic points for the current question, $p = \sum_{c \in C_{ch}} p_c \Rightarrow p = \sum_{c \in C_{ch}} \frac{p_{max}}{c_{max}} = \frac{p_{max}}{c_{max}} * c_{ch} = p_c * c_{ch}$

3.2.2 Scoring of the Penalty

Below we present our approach for scoring the penalty. We use the following designations:

- $a_{max} \in \mathbb{N}, a_{max} = |A|$ – number of options $a \in A$
- $Ch \subseteq A$ – set of *checked* options
- $ch = |Ch|, ch \in \mathbb{N}, ch \in [0, a_{max}]$ – number of checked options for the current question
- $b \in \mathbb{R}, b \in [0,1]$ – basic level of guessing for the current question, $b = \frac{c_{max}}{a_{max}}$
- $n \in \mathbb{R}, n \in [b,1]$ – measure, that shows the possibility, that user tries to guess the correct response by choosing too much options; we do not evaluate it in the cases, when $n <= b$,$n = \frac{ch}{a_{max}}$
- s – penalty for the guessing, $s = n - b \Rightarrow s \in [0, 1 - b]$
- $s_k \in [0, p_{max}]$ – the penalty, mapped to the maximal possible points.

  A mapping function will be: $f : s_k \rightarrow s$
  Given, $s_k \in [0, p_{max}]$ and $s \in [0, 1-b]$, then
  $f : s_k \rightarrow s = f : [0, 1-b] \rightarrow [0, p_{max}] \Rightarrow s_k = f(s) = s * \frac{p_{max}}{1-b} = (n-b) * \frac{p_{max}}{1-b}$
  The absolute score for the question is trivially determined as $T = f(p, s_k) = p - s_k$

4 EVALUATION

4.1 Synthetic Experiments

In the subsection we describe our experiments with synthetic data and compare the behavior of different methods. We consider all the questions to have the difficulty $d = 1$, then the maximal possible points $p_{max} = 1$ as well.

Table 1: Comparison of the proposed approach with other existing approaches.

Dich.	MTF	Morgan	Ripkey	Balanced
0	0.4	0	0	0

Example 1 (Case: 5 options, 2 correct, 5 marked)**.** *In the case the student chose all the options and should obtain zero points. However, we see that MTF method does not recognize this type of guessing and considers the questions to be answered partially correct, awarding the points for two correct options, that were marked.*

Table 2: Comparison of the proposed approach with other existing approaches.

Dich.	MTF	Morgan	Ripkey	Balanced
0	0.6	0	0	0

Example 2 (Case: 5 options, 2 correct, 0 marked)**.** *The situation is opposite to the previous: in the case the student chose none of the options. As we assume that question must have at least one correct option, in case of not choosing any options a student also should obtain zero points. However, we see that MTF method awards the points for three distractors, that were not marked. Although the situation is absurd, we faced it within real learning platforms, for example within several on-line courses of the Stanford University* [2].

Two examples below are trivial and the problem could be solved by adding the rules. However, the MTF scoring also suffers from skewness, when applied to MMQs, as it is shown below.

Table 3: Comparison of the proposed approach with other existing approaches.

Dich.	MTF	Morgan	Ripkey	Balanced
0	0.83	0.5	0.5	0.5

[2]http://online.stanford.edu/courses

Example 3 (Case: 6 options, 2 correct, 1 correct marked)**.** *This case proves, that the MTF method has a dependency from a number of correct and incorrect options. Thus, in a case of 6 options two of which are correct, a student is awarded 0.833 points for choosing only one correct option. In a case of 5 options two of which are correct, she would be awarded 0.80 points for the same. Moreover, if she choose only one incorrect option in a case of 6 alternatives, she obtains 0.5 points; in a case of 5 options she will be awarded 0.4 for the same.*

Thus, our experiments prove, that multiple-mark questions can not be scored properly with the algorithms, developed for multiple true-false items. However, the MTF scoring is the only existing approach of partial scoring that can be used in a case, when a question does not have any correct options.

Table 4: Comparison of the proposed approach with other existing approaches.

Dich.	MTF	Morgan	Ripkey	Balanced
0	0.5	0	0.5	0.5

Example 4 (Case: 4 options, 2 correct, 1 correct and 1 incorrect marked)**.** *This case illustrates the issues of using the Morgan algorithm. The Morgan algorithm deducts penalties for choosing the incorrect option, as well as the proposed approach. However, in that case we are facing the situation, that penalty has the same size, as the basic points, and the student is awarded zero. We consider the penalty to be needlessly high, especially because the penalty depends on the number of incorrect options. Thus, if the question has 3 incorrect options, choosing one of them would be fined on 0.33, and in case of 2 incorrect options, the penalty is 0.5. After recognizing behavior of the algorithm, students will mark only the options, they are sure in, because choosing an incorrect one may cost them a full amount of points, they collected with correct options.*

The next two examples show mainly the differences between the proposed approach and Ripkey algorithm. Namely, we show the situations, when Ripkey algorithm awards zero points, while we consider that it should award more.

Table 5: Comparison of the proposed approach with other existing approaches.

Dich.	MTF	Morgan	Ripkey	Balanced
0	0.75	0.5	0	0.5

Example 5 (Case: 4 options, 2 correct, 2 correct and 1 incorrect marked)**.** *In this case the student chose*

more options, than the number of correct ones, and according to the Ripkey, the answer should be awarded zero. Our claim is, that until the student have not chosen all the options, she could have some points. However, choosing three of four options could mean a try of guessing. Although in this case the student gets the full amount of basic points, she is fined on a half of them.

Table 6: Comparison of the proposed approach with other existing approaches.

Dich.	MTF	Morgan	Ripkey	Balanced
0	0.8	0.67	0	0.67

Example 6 (Case: 5 options, 2 correct, 2 correct and 1 incorrect marked)**.** *The example shows the disadvantage of the Ripkey algorithm more clear. It is not clear for the student, why she was awarded zero points, as she did not try to guess and answered partially correct.*

Table 7: Comparison of the proposed approach with other existing approaches.

Dich.	MTF	Morgan	Ripkey	Balanced
0	0.6	0.17	0.67	0.67

Example 7 (Case: 5 options, 3 correct, 2 correct and 1 incorrect marked)**.** *In that case balanced scoring and Ripkey algorithms behave the same, as none of them deducts a penalty.*

4.2 Real-life Evaluation

We have implemented the balanced scoring method within our e-learning system SlideWiki[3] (Khalili et al., 2012). For evaluation of our algorithm we used a lecture series on "Business Information Systems". We chose this course since it comprises a large number of definitions and descriptions, which are well suited for the creation of MMQs. In total we have created 130 questions. A course of 30 students was offered to prepare for the final examination using SlideWiki. Overall, the students made 287 attempts to complete the quiz and we collected all their answers (also unfinished assessments) for the evaluation. After collecting the answers, we implemented all discussed algorithms to score and compare the results, in particular with regard to the ranking and the mean score. The results are summarized in 1.

The study aimed to investigate two aspects of the proposed approach:

- How severe does the balanced scoring approach penalize?
- How does balanced scoring differ from Dichotomous scoring?

We answer the first question by comparing the scores calculated using all discussed algorithms for the same quiz (see 1, upper part). These two diagrams show, that on average the balanced scoring approach penalizes more severely than MTF scoring and less severely than other discussed approaches. We answer the second question by comparing the difference in student ranking. We rank all assessments based on the individual scores. That is, assessments with higher scores rank higher than assessments with lower scores and equal scores result in the same ranking. We compare the rankings of other approaches with the rankings calculated using the dichotomous scoring, since we consider the dichotomous scoring to be the ranking reference. The two lower diagrams in 1 show the results of this evaluation. They show, that the ranking of the balanced scoring approach is the closest to the dichotomous ranking when compared to the other algorithms.

5 CONCLUSIONS

In the paper we evaluate the existing approaches for scoring the multiple-mark questions and propose a new one. The proposed approach has a list of restrictions, however it has advantages when compare with the discussed approaches. One of the main advantages is that is based on the mathematical model, it does not suffer from the skewness, as it has the same formula for all cases. At the same time, the proposed approach recognizes the attempts to guess the correct answer, for example choosing all the possible options. When compare with the existing approaches, the advantages of the proposed algorithm could be summarized as follows:

- The approach allows to score both multiple-mark and conventional multiple-choice questions.
- The approach is based on the partial scoring concept.
- The algorithm can be easily implemented, it is pure mathematical.
- The score does not highly depend on the amount of correct and incorrect options.
- The value of the penalty is in balance with the possibility, that the student is trying to guess.
- Due to the balance, the results are clear for the students.

[3]http://slidewiki.aksw.org

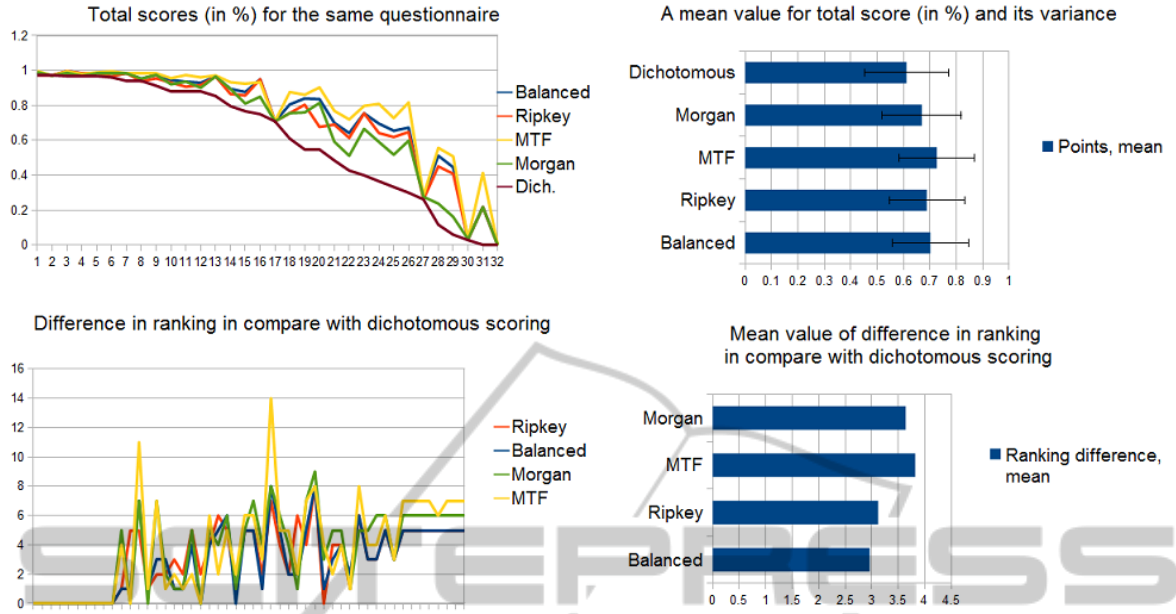

Figure 1: The statistics of the evaluation.

However, we suppose our algorithm to be optional together with other discussed approaches. This is due to the fact, that teachers create questions in their own manner and should be able to choose an appropriate method to score the results. Also, the different situations require different levels of severity, and the proposed approach might be too lenient.

REFERENCES

Bauer, D., Holzer, M., Kopp, V., and Fischer, M. R. (2011). Pick-N multiple choice-exams: a comparison of scoring algorithms. *Advances in health sciences education : theory and practice*, 16(2):211–21.

Ben-Simon, A., Budescu, D. V., and Nevo, B. (1997). A Comparative Study of Measures of Partial Knowledge in Multiple-Choice Tests. *Applied Psychological Measurement*, 21(1):65–88.

Cronbach, L. J. (1941). An experimental comparison of the multiple true-false and multiple multiple-choice tests. *Journal of Educational Psychology*, 32:533–543.

Dressel, P. and Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement*, 13(4):574–595.

Farthing, D., Jones, D., and McPhee, D. (1998). Permutational multiple-choice questions: an objective and efficient alternative to essay-type examination questions. *ACM SIGCSE Bulletin*, 30(3):81–85.

Frisbie, D. A. (1992). The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice*, 11(4):21–26.

Hohensinn, C. and Kubinger, K. D. (2011). Applying Item Response Theory Methods to Examine the Impact of Different Response Formats. *Educational and Psychological Measurement*, 71(4):732–746.

Itten, S. and Krebs, R. (1997). *Messqualitaet der verschiedenen MC-Itemtypen in den beiden Vorpruefungen des Medizinstudiums an der Universitaet Bern 1997/2*. Bern: IAWF.

Jiao, H., Liu, J., and Haynie, K. (2012). Comparison Between Dichotomous and Polytomous Scoring of Innovative Items in a Large-Scale Computerized Adaptive Test. *Educational and Psychological Measurement*, 72(3):493–509.

Khalili, A., Auer, S., Tarasowa, D., and Ermilov, I. (2012). Slidewiki: Elicitation and sharing of corporate knowledge using presentations. In *Proceedings of the EKAW 2012*, pages 302–316. Springer.

Morgan, M. (1979). MCQ: An interactive computer program for multiple-choice self-testing. *Biochemical Education*, 7(3):67–69.

Pomplun, M. and Omar, M. H. (1997). Multiple-Mark Items: An Alternative Objective Item Format? *Educational and Psychological Measurement*, 57(6):949–962.

Ripkey, D. and Case, S. (1996). A" new" item format for assessing aspects of clinical competence. *Academic Medicine*, 71(10):34–36.

Serlin, R. and Kaiser, H. (1978). A method for increasing the reliability of a short multiple-choice test. *Educational and Psychological Measurement*, 38(2):337–340.

Tsai, F. and Suen, H. (1993). A brief report on a comparison of six scoring methods for multiple true-false items. *Educational and psychological measurement*, 53(2):399–404.