The Scoring of Matching Questions Tests: A Closer Look

Antonín Jančařík and Yvona Kostelecká
Charles University in Prague, Faculty of Education, Prague, Czech Republic
antonin.jancarik@pedf.cuni.cz
yvona.kostelecka@pedf.cuni.cz

Abstract: Electronic testing has become a regular part of online courses. Most learning management systems offer a wide range of tools that can be used in electronic tests. With respect to time demands, the most efficient tools are those that allow automatic assessment. The presented paper focuses on one of these tools: matching questions in which one question can be paired with multiple response terms. The aim of the paper is to identify how the types of questions used in a test can affect student results on such tests expressed as test scores. The authors focus mainly on the issue of the possible increase in scores that can occur with the use of closed questions, when students, after selecting the answers to the questions they know the correct answers to, then guess the answers to the remaining questions (see Diamond and Evans, 1973, Ebel and Frisbie, 1986, Albanese, 1986). The authors show how the number of distractors (unused answers) included in a question influences the overall test score. The data on multiple-choice and alternative- response tests are well known, but not much is known about matching questions. Estimating formula scores for matching-question tests is important for determining the threshold at which students demonstrate they possess the required level of knowledge. Here the authors will compare the scores obtained for three types of closed questions: multiple choice, alternative response and matching questions. The analysis of matching assignments in this paper demonstrates that they are a useful tool for testing skills. However, this holds only if the assignment has at least two distractors. Then the informational value of this type of assignment is higher than that of multiple-choice assignments with three distractors. The results currently indicate that these types of assignment are not useful if the objective of the testing is to rank students or to distinguish between very good students – and this applies even if two distractors are used. In the case of such an objective, it is better to use multiple-choice assignments.

Keywords: testing, random score, test results, matching type, score formula, formula scoring

1. Introduction

A general objective of this paper is to determine how the use of closed test assignments and questions may influence student test scores, and from an analysis determine which types of test assignments are best and have the greatest discriminating power. We will estimate the scores students would attain and the probability of their attaining them if they know the answers to a certain number of questions and guess the answers to the rest (see Diamond and Evans, 1973, Ebel and Frisbie, 1986, Abu-Sayf, 1979 and Albanese, 1986). We will compare the obtainable scores for three types of closed questions - multiple choice, alternative response and matching questions - and for combinations of them within individual tests. The results of the calculations will be demonstrated on examples that show how the choice or use of a particular test influences a student's test score.

This study was motivated by the preparation and assessment of tests of Czech-language knowledge to be applied to the children of immigrants to the Czech Republic (see Kostelecka and Jancarik, 2014). The tests examined in the course of our research contained the various types of questions mentioned above. In subject areas studied in the research we combined various types of tests. In order to compare the results in individual subject areas it was necessary to distinguish the random score(s) in relation to the type of test used. We found that the issue of formula scoring in the case of multiple-choice and alternative-response (true/false) types of questions has been extensively discussed in the literature; however we were unable to find in the literature an analysis of formula scoring calculations for the matching type of question. Our calculations, which are presented below, show that matching questions have different attributes from the other two question types. Most notably, the score formula for this type of question is not a linear function, which means that it is possible to change the properties of a test using matching questions, particularly the area in which the test possesses a best discriminating power. We will demonstrate this aspect of matching questions using the example of the situation that motivated this research, namely, the need to create language tests that distinguish between students on the basis of a 60% level of knowledge of the material tested.

Reference this paper as Jančařík A and Kostelecká Y. "The Scoring of Matching Questions Tests: A Closer Look" *The Electronic Journal of e-Learning Volume 13 Issue 4 2015*, (pp270-276) available online at www.ejel.org

2. Multiple-choice questions

There is currently a wide range of programmes that can be used to create matching-type questions and enable their use both on websites and in almost every type of e-learning system. One frequently used programme for creating this type of test is Hot PotatoesTM (see Figure 1), for which there also exists a plugin for integration with the Moodle Language Management System (LMS).

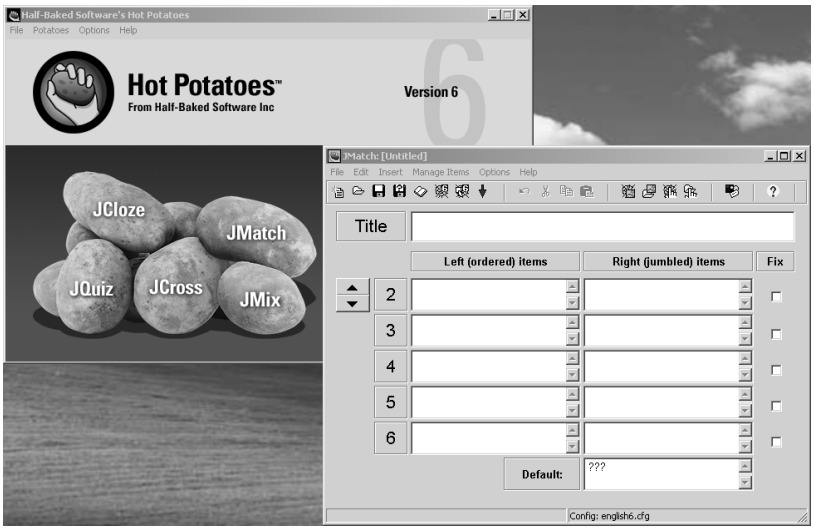

Figure 1: Hot Potatoes™

Individual programs differ, of course, in terms of the number of options they offer and their visual presentation (see Figure 2). Usually, however, they offer the user the option to choose the number of questions and possible answers.

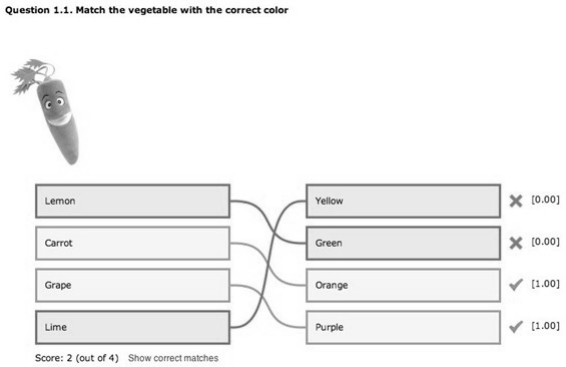

Figure 2: Matching-type question in LMS Fronter

3. Calculating probabilities

To calculate the possible scores that students can obtain and the probability of students obtaining those scores we used classic probability and combinatorial methods (see Charter, 2000). We shall assume in reference to all the calculations made in this paper that the student always knows the answers to a pre-determined number of questions and that he or she guesses the answers to the remaining questions by using each of the other responses just once. Programs automatically ensure that one answer cannot be applied to more than one question. The number of questions and the number of response options that are included in a test have an effect on the total random score that a student can attain. In this analysis we are focusing on tests in which each question comprises five sub-questions. This is the same number that was used in the test studied in our previous research. Moreover, it is easy to fix the total number of points awardable using this number of sub-questions and to create combinations of questions for tests of different length.

For each question and sub-question we examined the use of between five and seven response options. These represent three different approaches to formulating a test assignment that asks a student to match five items and to do by choosing among:

- five response options (the 5-5 type),
- six response options (the 5-6 type) and
- seven response options (the 5-7 type).

The results we compare here are of the scores students would probably obtain using the above-mentioned types of test tasks and of the scores they would obtain if we used multiple-choice or alternative-response questions.

To calculate the probability of students obtaining a certain score (see Arratia and Tavare, 1992, Pitman, 1997) we used rencontres numbers $\mathrm{F(k,n)}$ as described, for instance, by Riedel (2006), where $\mathrm{F(k,n)}$ is the number of permutations of an *n*-element set that keeps *k* elements fixed.

$$F(k,n) = n! \sum_{j=k}^{n} \frac{-1^{j-k}}{(j-k)!\,k!}$$

3.1 The 5-5 type of matching question

In the case of the 5-5 type of matching question, the student is presented with five response options (without using distractors) and has to correctly match them to five lexical items. Table 1 presents the calculated probabilities of obtaining individual scores. The rows give the number of questions the student would answer correctly, while the columns give the probability of the given score being attained. We are interested in learning, for instance, what the probability is that a student who knows the answer to fewer than three questions will ultimately obtain three or more points in total in this kind of assignment (and will thus obtain a passing score of 60% or more). Table 1 shows that the probability of students who know the answer to just two questions getting a score of three or more points is greater than 50% in this type of assignment (there is a 50% likelihood that the student will obtain three points, and a 17% likelihood that the student will obtain as many as five points for this assignment).

Table 1: The 5-5 type of matching question: the number of answers a student knows – the total score

	0	1	2	3	4	5
0	37%	38%	17%	8%	0%	1%
1		38%	33%	25%	0%	4%
2			*33%*	*50%*	*0%*	*17%*
3				50%	0%	50%
4					0%	100%
5						100%

3.2 The 5-6 type of matching question

Table 2 presents the score and probability calculations for a matching question that uses one distractor (the 5-6 type). This type of of assignment is of more informational value than than a multiple-choice question with three distractors. However, students who know the answer to just two questions here still have more than a 50% probability of obtaining at least the required score of 3 points and and thereby passing this assignment.

Table 2: The 5-6 type of matching assignment: 5 questions and 6 matching options (1 distractor)

	0	1	2	3	4	5
0	43%	37%	15%	4%	1%	0%
1		44%	37%	15%	3%	1%
2			46%	38%	13%	4%
3				50%	33%	17%
4					50%	50%
5						100%

3.3 The 5-7 type of matching question

Table 3 shows the probabilities of different scores being obtained in a matching question with two distractors (the 5-7 type) in relation to the number of answers a student truly knows. In this type of assignment, the probability that a student who knows two correct answers will obtain three points is less than 50%. Among the assignments studies here this type will be the one best suited to testing the language skills of immigrant students in the proposed diagnostic test.

Table 3: The 5-7 type of matching assignment: 5 questions and 7 matching options (2 distractors)

	0	1	2	3	4	5
0	48%	36%	13%	3%	0%	0%
1		50%	36%	12%	2%	0%
2			*51%*	*38%*	*10%*	*2%*
3				58%	33%	8%
4					67%	33%
5						100%

3.4 A comparison of tests

Tables 4 and 5 present comparisons of all the above-mentioned test assignments. The results for the tests that use multiple-choice (M-C) and alternative-response (T-F) questions are also included in these tables for comparison. Table 4 calculates average test scores for each type of test in relation to the number of correct answers a student knows. The results presented in the figures indicate that the most informative type of assignment for diagnostic testing is the matching question test with two distractors. Table 5 shows the probability that a student who knows the correct answer to two or fewer questions will attain the required minimum of three points to pass the assignment. This threshold corresponds to the requirements of the language test that the initial calculations were prepared for. The results indicate that the alternative-response type of assignment (true/false) is not appropriate because it is of little informational value. The best and most informative type of assignment is the matching question with two distractors. Pre-testing moreover showed that this type of test assignment appeals to students and is easy to understand.

Table 4: Average scores for the different types of test assignment

	T-F	M-C	5-5	5-6	5-7
0	2.5	1.25	1.0	0.8	0.7
1	3.0	2.00	2.0	1.8	1.7
2	3.5	2.75	3.0	2.8	2.6
3	4.0	3.50	4.0	3.7	3.5
4	4.5	4.25	5.0	4.5	4.3
5	5.0	5.00	5.0	5.0	5.0

Table 5: The probability a student can successfully pass the assignment even if she/he has less than the required amount of knowledge to pass

	T-F	M-C	5-5	5-6	5-7
0	50%	10%	9%	5%	3%
1	69%	26%	29%	19%	14%
2	88%	58%	67%	54%	49%

4. Calculating probabilities for a combination of questions

The calculations show that the score formula for the matching-type /assignment is not (unlike the formula scoring for other types of questions studied) a linear function (cf. Ridel, 2006). As a result, the matching question is better at accurately measuring a student's skills on some levels than other types of questions are. On the other hand, when multiple questions are used the overall resulted is influenced by how the student's knowledge is distributed between different questions. Table 6 shows the situation where a student is presented with two matching-type questions and knows the answers to six sub-questions. The average attained score is divided according to how the student's knowledge of the answers to sub-questions is divided.

The results show that students whose knowledge is distributed evenly among the questions have a slight advantage. This fact needs to be taken into account in the development and assessment of tests.

Table 6: Average score for two questions of the same type based on a distribution of six correct answers

	5-5	5-6	5-7
5-1	7	6.8	6.7
4-2	8	7.3	6.9
3-3	8	7.4	7

5. Score formula and formula scoring for Multiple-Choice test

The basic objection to the use of tests with closed questions is that students often get part of their score by just guessing the answers to questions they cannot answer. If a test contains multiple-choice items, the typical student's strategy will be to answer those questions they know the answer to and then guess the answers to the rest of the questions. The results that we get are distorted and cannot be compared to the results of other tests. If, for example, we make two tests with the goal of comparing pupils' knowledge, where one test consists of open questions and the other is multiple choice with just two items per question, the score on the second test will naturally be much higher because most of the answers will have been guessed with a fifty percent probability of guessing the right answer.

There are several ways to rectification of the gained data. Each method has its advantages and drawbacks and which one is used will depend on the particular test assignment and the goal we want to achieve. There are three basic goals in all testing:

- to minimise wherever possible the number of questions in which students can make random guesses
- to minimise the differences (in scores/results) of students who know the same number of correct answers
- to ensure the comparability of results attained using different kinds of tests

The literature describes in detail some tools that can be used to minimise the number of answers a student can just guess with respect to one of the above stated goals (cf. Budescu and Bo, 2014, Farrell and Farrell, 2014). These tools include advice on how to select distractors, how to pose questions, and may recommend deducting points for incorrect answers, or they may give advice on how to estimate the number of guessed responses on the basis of the number of wrong answers. However, none of these options solves the problem entirely. For example, a pupil could rule out some of the response options to a multiple-choice question, thereby reducing the actual number of options from which to attempt to guess the right answer; the chances

of the student's success are thereby increased. The usual approach to estimating the number of questions to which a student did not know the answer in a multiple-choice test is based on the number of incorrect answers the student chose. This calculation is based on the idea that a student scores by chance in about one nth of the number of guessed answers (where *n* is the number of items to choose from). Thus, if a student answered 3 questions incorrectly on a test in which question was accompanied by four response items to choose from, we presume the student had been guessing four times and one of the guessed answers was correct. Thus the following formula could be used:

$$FS(C) = C - \frac{W}{(n-1)}$$

where $FS(C)$ is the formula scoring, C the number of correct answers and W the number of wrong answers. The table 7 presents the converted values for a test with five questions offering different numbers of response items to choose from (two, three, four and five items).

Table 7: Standard formula scoring for a multiple-choice test made up of five questions with different numbers of response items (from 2 to 5) to choose from

C/n	2	3	4	5
0	-5.00	-2.50	-1.67	-1.25
1	-3.00	-1.00	-0.33	0.00
2	-1.00	0.50	1.00	1.25
3	1.00	2.00	2.33	2.50
4	3.00	3.50	3.67	3.75
5	5.00	5.00	5.00	5.00

However, this method of conversion is not suitable for comparing results from different types of tests as it introduces negative values to the test results and in some cases significantly changes the range of possible test values. Moreover, the formula scoring presented above cannot easily be extended for use with matching-type tests.

For this reason the authors of this paper introduce a new approach based on the formula

$$SF\ (C) = \frac{\sum_{i=1}^{C} i \cdot P(C,i)}{\sum_{i=1}^{C} P(C,i)}$$

where C stands for the number of correct answers and $P(C,i)$ stands for the probability that a student will attain C correct answers when s/he knows i correct answers and guesses the remaining answers. This formula is based on [the idea ofthe mean value of a random variable and calculates the value from which the required number of points is attained if the student is guessing randomly.

Table 8: Score formula for a multiple-choice test made up of five questions with different numbers of response items (from 2 to 5) to choose from

C/n	2	3	4	5
0	0.00	0.00	0.00	0.00
1	0.29	0.38	0.44	0.50
2	0.73	0.97	1.14	1.27
3	1.43	1.87	2.14	2.31
4	2.53	3.08	3.36	3.51
5	4.10	4.51	4.67	4.75

Table 9: Score formula for a multiple-choice test made up of ten questions with different numbers of response items (from 2 to 5) to choose from

C/n	2	3	4	5
0	0.00	0.00	0.00	0.00
1	0.17	0.23	0.29	0.33
2	0.39	0.56	0.70	0.83
3	0.69	1.02	1.30	1.53
4	1.11	1.67	2.12	2.45
5	1.71	2.57	3.17	3.56
6	2.55	3.72	4.38	4.76
7	3.73	5.06	5.68	6.00
8	5.25	6.51	7.00	7.25
9	7.05	8.00	8.33	8.50
10	9.01	9.50	9.67	9.75

Tables 8 and 9 show how the converted test results change depending on the number of response items. We can see that the range of results is close to the range of results on the original tests results. Unlike the above-described method usually used, this method keeps the minimum value but also decreases the maximum value because the maximum number of points can be attained not only thanks to knowledge, but also to random guessing. The presented function is not, unlike the previous one, a linear function.

5.1 Example of use

The differences between both functions after the correction of test results can be demonstrated by comparing results in a test of language skills of migrant children at the B2 level in writing and listening (see Kostelecka and Jancarik, 2014). In the written test, pupils produced written answers to open questions, in the listening test true-false question were used. Therefore we compared a test with open questions to a test highly prone to random error. The results of this comparison are presented in Table 10. The comparison was carried out both for a complete set of data and for data from which we excluded tests on which pupils scored either the minimum or the maximum number of points (the biggest difference between the conversion functions is at these two extremes). The comparison shows that despite the significant initial difference in the point values, the pupils' skills in both studied areas were very similar.

Table 10: A comparison of calculations before and after subtracting randomly attained calculations using both methods described above

	Ø Writing	Ø Listening	Ø Listening SF	Ø Listening FS
Full sets of data	3.46	6.98	4.12	3.72
Data without extremes	4.80	7.26	4.41	4.30

Table 11: A comparison of calculations before and after subtracting randomly attained calculations using both methods described above

	Ø |Writing - Listening|	Ø |Writing - Listening SF|	Ø |Writing - Listening FS|
Full sets of data	3.73	2.48	3.08
Data without extremes	2.70	2.29	2.95

This means that the conversion functions approximated the data of both groups. What is very interesting is to compare the values before and after conversion for individual students. In Table 11 you will find the 'average distance' between the results of individual students in both skill areas studied. The table clearly shows that the

method we propose reflects the overall shift and approximation of results even at the level of the individual student, which is not true in the case of the standard FS method, where this approximation (in lesser degree) can only be observed for the full data set.

6. Determining the score formula for matching-type tests

Table 12: Score formula for each type of test and the attained score

	0	1	2	3	4	5
5-7	0.00	0.58	1.38	2.36	3.47	4.62
5-6	0.00	0.54	1.32	2.25	3.28	4.42
5-5	0.00	0.50	1.19	2.07		4.00
T-F	0.00	0.29	0.73	1.43	2.53	4.10
MC 3	0.00	0.38	0.97	1.87	3.08	4.51
MC 4	0.00	0.44	1.14	2.14	3.36	4.67
MC 5	0.00	0.50	1.27	2.31	3.51	4.75

The method of determining score formula presented above, based on probability of the different resultscan be modified for matching-type tests. The following table (Table 12) presents the corresponding values of the function for all the three types of tests studied here. For easy comparison it also contains data on multiple-choice tests.

7. Conclusion

The goal of this paper was to describe the basic properties of a matching-type test. The matching-type test has significant potential and is a tool particularly well suited to tests that seek to assess the level of knowledge a student has attained. For example, the matching-type test with two distractors is very good at distinguishing knowledge levels measured against a 60% passing score.

The results currently indicate that these types of assignment are not useful if the objective is to rank students or to distinguish between very good students – and this applies even if two distractors are used.

The paper introduces two methods of rectification of data that are obtained from tests made up of closed questions. The rectification calculations make it possible to compare scores attained in different types of test because they allow the score values to be 'purged' of random score increases that can occur in relation to the type of test used. The paper compares two methods used for multiple-choice tests and introduces how one of the methods can be modified to be used with matching-type tests. This paper presents the score values of this scoring formula for the 5-5, 5-6 and 5-7 types of test and introduce a method that can be used to calculate these values also for other types of test.

Acknowledgements

This article was financially supported by the Czech Science Foundation within the project *Integration of the children of non-nationals into the Czech elementary schools*, registration number: 13-32373S

References

Abu-Sayf, F. K. (1979) "The scoring of multiple-choice tests: A closer look", *Educational Technology*, Vol. 19, pp. 5-15.

Albanese, M. A. (1986) "The correction for guessing: A further analysis of Angoff and Schrader", *Journal of Educational Measurement*, Vol. 23, pp. 225-235.

Arratia, R. and Tavare, S. (1992) "The cycle structure of random permutations", *Annals of Probability*, Vol. 20, pp. 1567-1591.

Budescu, D.V. and Bo, Y. (2014) "Analyzing Test-Taking Behavior: Decision Theory Meets Psychometric Theory", *Psychometrika*, pp. 1-18.

Charter, R. A. (2000) "Determining random responding to objective tests", *Journal of Psychoeducational Assessment*, Vol. 18, pp. 308-315.

Diamond, J. and Evans, W. (1973) "The correction for guessing", *Review of Educational Research*, Vol. 43, pp. 181-191.
Ebel, R. L. and Frisbie, D. A. (1986) *Essentials of educational measurement* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.
Farrell, G. and Farrell, V. (2014) Improving learning through interactive multiple choice questions with confidence measurement, *Proceedings of the IASTED International Conference on Computers and Advanced Technology in Education,* CATE 2014, pp. 1-8.
Kostelecká, Y. and Jančařík A. (2014) "The process of Czech language acquisition by foreign pupil at lower secondary school", *Journal of Efficiency and Responsibility in Education and Science*, Vol. 7, No. 1, pp. 8-13.
Pitman, J. (1997) "Some probabilistic aspects of set partitions", *American Mathematical Monthly*, Vol. 104, pp. 201-209.
Ridel, M. R. (2006) "The statistics of random permutations", [online], http://www.oocities.org/markoriedelde/papers/randperms.pdf