Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2014, Article ID 240263, 8 pages
http://dx.doi.org/10.1155/2014/240263

Research Article

New Indices for Refining Multiple Choice Questions

Mariano Amo-Salas,[1] María del Mar Arroyo-Jimenez,[2,3] David Bustos-Escribano,[3] Eva Fairén-Jiménez,[3] and Jesús López-Fidalgo[1]

[1]*Department of Mathematics, Institute of Applied Mathematics in Science and Engineering, University of Castilla-La Mancha, 13071 Ciudad Real, Spain*
[2]*Department of Medical Sciences, University of Castilla-La Mancha, 13071 Ciudad Real, Spain*
[3]*Medical Education Unit, University of Castilla-La Mancha, 13071 Ciudad Real, Spain*

Correspondence should be addressed to Jesús López-Fidalgo; jesus.lopezfidalgo@uclm.es

Received 8 September 2014; Accepted 7 December 2014; Published 23 December 2014

Academic Editor: Chin-Shang Li

Multiple choice questions (MCQs) are one of the most popular tools to evaluate learning and knowledge in higher education. Nowadays, there are a few indices to measure reliability and validity of these questions, for instance, to check the difficulty of a particular question (item) or the ability to discriminate from less to more knowledge. In this work two new indices have been constructed: (i) the no answer index measures the relationship between the number of errors and the number of no answers; (ii) the homogeneity index measures homogeneity of the wrong responses (distractors). The indices are based on the lack-of-fit statistic, whose distribution is approximated by a chi-square distribution for a large number of errors. An algorithm combining several traditional and new indices has been developed to refine continuously a database of MCQs. The final objective of this work is the classification of MCQs from a large database of items in order to produce an automated-supervised system of generating tests with specific characteristics, such as more or less difficulty or capacity of discriminating knowledge of the topic.

1. Introduction

Tests based on multiple choice questions (MCQs) are widely used for evaluation. These tests are basically designed to assess learning and knowledge. Nevertheless, tests may be built carefully to asses other capacities as clinical reasoning. There are some recommendations to take all this into account [1–3]. It is widely accepted that well-constructed MCQs are time consuming and difficult [4, 5], which justifies a careful review of each of the items. The main advantage of this methodology is to provide feedback to both the students and the professors.

MCQs are items with a *stem* (starting part of the item, e.g., a question or a statement to be completed) and a set of k possible *responses*, generally ranging from 3 to 5. The only correct response is usually called the *key* and the incorrect responses are called *distractors*. The students have to select just one response or none. The mark is 1 if the answer is correct, 0 if none of the responses have been chosen and there is a penalty of $1/(k-1)$ for each failure. Thus, in this work we are considering a correction for guessing. This penalty is an unbiased estimate of what a student can get when answering randomly if there is no penalty. A negative number for the final mark is theoretically possible, but this rarely happens in practice with a sufficient number of items, which is crucial for this type of test.

We focus this work on this type of MCQs due to practical reasons. In Spain, after obtaining the B.M. degree (six years), all the graduates have to pass a national competitive examination based on MCQs to access a specialty in medicine. After passing the examination, all the graduates are ranked and they can choose specialty from different offers to spend a training period of 3–5 years in a medical center. The national competitive examination to access a specialty in medicine consists of 225 multiple choice questions with five options of which only one is correct and 10 questions in reserve in case formulation problems or errors are detected (235 in total). The mark is 1 if the answer is correct and 0 if none of the responses have been chosen and there is a penalty of 1/4 for each failure. As a matter of fact this kind of test is used in

almost all the faculties of medicine in Spain to get the students used to it. Moreover, this is the type of MCQs generally used in higher education in Spain.

One of the main characteristics of these items is the existence of indices to analyze their reliability and validity, for instance, the difficulty or discrimination index. These indices allow the categorization of these items based on the obtained answers. Another utility of these indices is to detect mistakes in the items providing a tool to improve the item for future use. They can also be used to investigate why more failures than usual are observed in a particular item. The difficulty of a particular item may be caused by reasons intrinsic to the item (e.g., a complex concept) or because the key or the distractors lead to the failure of the student. Most of the poor designed items are characterized by the following: (i) the item not succeeding in assessing the main objective, (ii) existence of clues for the right answer, and (iii) the text of the stem or the responses being ambiguous. The aim of the distractors is to look like plausible solutions to the problem for those students who do not achieve the objective assessed in the item. At the same time the distractors have to be not plausible for the students reaching the objective evaluated by the item. For these students just the correct answer has to be plausible.

There are some indicators to identify weak and strong groups or to measure the difficulty and discrimination capacity of items and tests. As far as the authors know the literature on this topic does not consider any measure neither of the homogeneity of the responses nor of the rate of the "no answers" [6]. There exist a group of techniques based on fuzzy approach, based on more complicated ordering of results enabling the student to explicitly describe his/her degree of confidence in each possible answer [7].

The aim of this paper is to provide two new indices to measure the relationship between the number of errors and the number of no answers as well as the homogeneity of the responses of an item. As a matter of fact the justification of the penalty described above is strongly based on the homogeneity of the distractors and any violation of this hypothesis makes the use of the penalty inadequate. The indices provided here will help in checking this intrinsically in order to get a suitable test.

Finally, in this paper, a joint analysis of different indices is developed in order to obtain a procedure of classification of MCQs to detect the items that should be revised. In this sense the algorithm works as a security system.

2. Materials and Methods

2.1. Difficulty and Discrimination Indices. Difficulty and discrimination indices are classic in the analysis of MCQs and they have been widely treated in the literature [8, 9].

The *difficulty index* (P) is defined as the proportion of correct answers among the students who did the test:

$$P = \frac{N_1}{N}, \tag{1}$$

where N is the number of students who performed the test and N_1 is the number of students who answered correctly the item. Thus, it is within the interval $[0, 1]$.

This index may be used to compare the difficulty of a particular item with the global difficulty of the test. Thus, this index may be used to check the homogeneity of the test in the sense of difficulty.

The *discrimination index* (D) measures the capacity of an item to distinguish between different levels of knowledge of the students. In order to compute this index the students tests have to be sorted from lower to higher scores. Then a group with the lowest scores (lower group) and another group with the highest scores (upper group) are built. The size of these groups varies according to the literature, but it is usually around 30% of the total number of students. The most frequent size in the literature is 27%, for example, [10]. Other sizes may be found, for instance, in Tristrán [11]. The definition of the index is then

$$D = P_u - P_l, \tag{2}$$

where P_u is the proportion of the students in the upper group who answered correctly the item and P_l is the proportion of the students in the lower group who answered correctly the item. The values of the index are in the interval $[-1, 1]$, where 1 means maximum discrimination and 0 means minimum discrimination. Negative values of the index mean that the students of the upper group failed with this item more than the students of the lower group, which is contradictory with what is expected.

Although it is expected that difficult items will discriminate better than easy items, this is not always the case and the combination of both indices provides an interesting tool to check possible incoherencies. Both indices are based just on the correct answer, but the rest of the responses play an important role as well. The homogeneity index given in this paper considers all the responses.

2.2. Homogeneity Index of the Distractors. A new index is defined to measure the homogeneity of the distractors in a MCQ. Thus, this index measures whether the number of wrong answers is equally distributed among all the responses, justifying the use of the traditional penalty. If there is some response with very low frequency, this means that for the students it is too obvious that this response is wrong and the students who chose this distractor are penalized in the same quantity compared to those who chose a more feasible distractor. On the contrary, if there is some distractor with very high frequency, this means that this response may be ambiguous and leads the students to a wrong interpretation.

The importance of this index comes from the penalty the student receives from a wrong answer. This penalty is based on the hypothesis that all the responses have the same difficulty and therefore the same chance to be chosen at random. Then a person choosing an answer randomly may have more probability to succeed than a person who studies the topic and is confused by an unclear interpretation of one of the responses. A higher frequency may be considered more unfair for the student than a lower one.

The index given here is based on the lack-of-fit test. Let N be again the number of students and k the number of responses for each item. Let $N = N_0 + N_1 + E$, where N_0 is

Table 1: Critical values for coefficient H for small number of errors and a significance level of 2.5%.

$k\backslash E$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
5	3	6	9	6	8.6	7.33	8.43	9	9.22	8.4	9	8.67	9.46	8.29	8.73	9	9.11	9.11	9
4	2	4	6	8	5.2	7	5.43	7	6	6.2	7.82	6.5	7.54	7					
3	1	2	3	4	5	6	3.58	4.5	5.44	3.6									

the number of people marking none of the responses, N_1 is the number of successes, and E is the number of failures (errors). Moreover $E = N_2 + \cdots + N_k$, where N_i, $i = 2, \ldots, k$, are the numbers of students choosing each of the $k - 1$ distractors.

The later numbers follow a multinomial distribution of size E:

$$(N_2, \ldots, N_k) \equiv M(E, p_2, \ldots, p_k), \tag{3}$$

where $p_i = N_i/N$ is the proportion of subjects selecting response i and $p_2 + \cdots + p_k = 1$.

To apply the traditional penalty, the optimal situation is that all the responses would have the same level of difficulty and therefore the frequencies should be similar. The following is a typical lack-of-fit hypotheses test:

$$\begin{aligned} H_0 &: p_2 = \cdots = p_k, \\ H_1 &: p_i \neq p_j \quad \text{for some } i, j. \end{aligned} \tag{4}$$

The explicit formula for the index is the test statistic:

$$H = \sum_{i=2}^{k} \frac{\left(N_i - E/(k-1)\right)^2}{E/(k-1)}. \tag{5}$$

The probability distribution of this statistic is approximated by a chi-squared distribution with $k - 2$ degrees of freedom. The values of this distribution can be found in any text book of basic statistics or in any statistical software, including Excel (`=CHIIN (probability; degrees of freedom)`). This approximation is good enough if most of the expected frequencies are greater than or equal to 5 and none of them is less than 1.5 [12]. Index H may vanish in two very different cases. On the one hand, if there is perfect homogeneity, then $N_i = E/(k-1)$ for every i. On the other hand, $H = 0$ if there are no errors. In the latter case the index should not be applied while the first case means that there is not any clear objection against homogeneity. Table 1 gives critical values, for a significance level of 2.5%, for low numbers of errors computed with 200,000 simulations for each one. For example, if $k = 5$ and $E > 19$, the critical value is 9.348 for a significance level of 2.5%. Notice that if the number of errors is too small, the H index is still coherent. For instance, if $k = 5$ and $E = 1$, then the observed index is 3 and the critical number is 3 and therefore there is no evidence of nonhomogeneity. If $E = 2$, they may be distributed in two distractors ($H = 2 < 6$) or concentrated at the same distractor with $H = 6$, which is the critical value and therefore there is no evidence yet to assert lack of homogeneity. Later on, we will explain why we use here 2.5% as significance level instead of the traditional 5%.

Table 2: Critical values for coefficient G for small number of errors and a significance level of 2.5%.

$N_0 + E/p_0$	0.1	0.2	0.3	0.4	0.5	$N_0 + E/p_0$	0.1
1	3.00	2.00	1.53	1.22	1.00	26	2.22
2	1.89	2.83	2.16	1.73	1.41	27	2.76
3	3.27	2.02	2.65	2.12	1.73	28	2.65
4	2.67	2.75	1.96	2.45	2.00	29	2.54
5	2.24	2.24	2.44	1.83	2.24	30	2.43
6	3.27	2.86	1.96	2.17	2.45	31	2.33
7	2.90	2.46	2.39	2.47	1.89	32	2.24
8	2.59	2.12	2.01	2.02	2.12	33	2.73
9	2.33	2.67	2.40	2.31	2.33	34	2.63
10	3.16	2.37	2.07	1.94	1.90	35	2.54
11	2.91	2.11	2.43	2.22		36	2.44
12	2.69	2.59	2.14	2.47		37	2.36
13	2.50	2.36	2.48	2.15		38	2.27
14	2.32	2.14	2.22			39	2.72
15	3.01	2.58	2.54			40	2.64
16	2.83	2.38	2.29			41	2.55
17	2.67	2.18	2.59			42	2.47
18	2.51	2.59				43	2.39
19	2.37	2.41				44	2.31
20	2.24	2.24				45	2.24
21	2.84	2.62				46	2.65
22	2.70	2.45				47	2.57
23	2.57	2.29				48	2.50
24	2.45	2.65				49	2.43
25	2.33	2.50				50	2.36

The value of H depends just on the differences between the number of observed errors in each response and the total number of errors divided by the number of distractors. The value of H increases when the number of subjects selecting a wrong answer is far from the expected value.

2.3. No Answer Index. Nevertheless, none of the indices considered so far takes into account the "no answer." We believe this is crucial to evaluate the suitability of an item because there is an important difference between an item where there is a large number of "no answer" and an item where there is a large number of errors. The first may mean that the stem of the question is ambiguous and the students do not understand the item. A large number of errors may be

due to a distractor very similar to the correct answer. Using again the lack-of-fit test for one proportion,

$$H_0: p = p_0,$$
$$H_1: p \neq p_0, \tag{6}$$

where p is the proportion of the errors among the "non-correct" answers and p_0 is a reference proportion. This proportion may be chosen from the whole set of items. This proportion of reference is sequentially fitted once a new validated item enters the database. The probability distribution of the statistic given by (7) is approximated by the standard normal

$$G = \frac{E/(E + N_0) - p_0}{\sqrt{p_0(1 - p_0)/n}}. \tag{7}$$

This is a bilateral test and the null hypothesis is rejected for large values of the absolute value of the statistic. For example, the critical value for a significant level of 2.5% is 2.24; thus if $|G| < 2.24$ the proportion is fine. The approximation is good for $np_0 > 5$ and $n(1 - p_0) > 5$; otherwise the critical values of Table 2 should be used. The meaning of a rejection depends on the sign of the statistic. Thus, if $|G|$ is greater than the critical value and $G > 0$, then there are too many errors; otherwise ($G < 0$) there are too many no answers.

2.4. Algorithm for Analyzing MCQs. The four indices considered in this paper, no answers (G), homogeneity of the distractors (H), difficulty of an item (P), and discrimination between the strong and the weak groups (D), are combined in order to offer a procedure to classify each item:

(i) Step 1: classification of the difficulty of an item using index P;

(ii) Step 2: discrimination capacity of an item using index D: if the index is within the range, then the item discriminates between the lower and the upper groups;

(iii) Step 3: if the index value of index G shows an appropriate proportion of "no answers" for the corresponding significant level, go to Step 4; otherwise go to Step 7;

(iv) Step 4: homogeneity of the wrong answers using coefficient H: the coefficient H gives the degree of homogeneity for a particular significance level; if an item has less than $5(k - 1)$ errors the chi-squared approximation should not be used and the exact multivariate distribution has to be considered instead (Table 1). If the item is homogeneous, go to Step 5; otherwise move to Step 6;

(v) Step 5: classify the item according to the indices;

(vi) Step 6: review the distractors causing nonhomogeneity;

(vii) Step 7: analysis of G: if the value of G is less than 0, the item should be reviewed very carefully; otherwise go to Step 8;

Table 3: Categorization for the difficulty and discrimination indices.

Index	Classification	Lower	Upper
Difficulty index (P)	Easy	0.75	1
	Moderate	0.25	0.75
	Difficult	0	0.25
Discrimination index (D)	High	0.3	1
	Moderate	0.2	0.3
	Low	–1	0.2

(viii) Step 8: analysis of homogeneity (H): if the wrong answers are homogeneous, the stem should be revised; if they are not, the corresponding distractor causing nonhomogeneity must be checked.

Figure 1 shows a scheme of the algorithm. From the algorithm, an item is being reviewed if at least one of the indices H or G is large. Thus, the significant level for each index, say α, should be adjusted to produce a global significant level of 0.05, which is the probability of the complementary of "failing in rejecting the null hypothesis at least in one of the tests when it is true." This means $0.05 = 1 - (1 - \alpha)^2$ and then $\alpha \approx 0.025$. Bonferroni's method gives exactly this number. This is the reason of using 2.5% as significant level for both indices.

This algorithm works as a security system where the alarms sound for particular items that should be revised in order to detect inadequate responses or questions. For instance, the lack of homogeneity detected for an item may mean it is badly designed or else it is just a false alarm because the heterogeneity of the distractors was intentioned by the professor.

3. Results

The algorithm was applied to the first year exams in the Faculty of Medicine. The students performed 5 progress tests of each of the 10 courses of the first year. Each paper had 10 MCQs on average with 5 possible responses for each one. Thus, the algorithm was applied to a total of 500 items answered by a slightly less than 50 students on average.

Table 3 shows the ranges of the P and D indices used for determining whether an item satisfies the criteria of difficulty and discrimination. For index G, the reference value (p_0) considered was 0.44; this value is based on the whole set of items studied. Therefore, for each item, the expected number of errors is a bit less than the number of no answers.

Table 4 shows examples of some items with the corresponding indices. Some comments on the examples follow ahead in order to show the utility and interpretation of the indices developed in this paper.

(i) Item P4 is nonhomogeneous since a majority of the students chose distractor C and the number of errors is high. Moreover, this item does not discriminate well.

(ii) Item P27 is a good example of a difficult question; it is homogeneous and discriminates quite well and the value of index G is fine.

Table 4: A sample of the results for different types of items.

Item	Correct response	Distribution of answers						Correct answers	Errors	Correct answers	Correct answers	P	H	D	G
		A	B	C	D	E	NA		E	GF	GD				
P4	B	2	13	15	1	2	8	13	20	3	4	0.32	26.8	−0.09	2.91
P27	D	1	5	5	7	9	16	7	20	5	0	0.16	6.4	0.45	1.38
P45	D	0	0	1	32	7	3	32	8	11	4	0.74	17	0.64	1.91
P46	A	32	0	1	7	0	3	32	8	11	8	0.74	17	0.27	1.91
P83	D	0	1	3	23	5	14	23	9	11	1	0.5	6.56	0.83	−0.48
P123	E	0	3	3	1	25	18	25	7	8	6	0.5	3.86	0.15	−1.62
P257	B	5	28	2	0	4	11	28	11	11	2	0.56	5.36	0.69	0.56
P259	B	2	30	2	7	0	9	30	11	12	1	0.6	9.73	0.85	0.98
P389	C	33	0	5	0	0	12	5	33	2	0	0.1	99	0.15	3.95
P404	A	9	5	2	5	0	29	9	12	5	1	0.18	6	0.31	−1.91

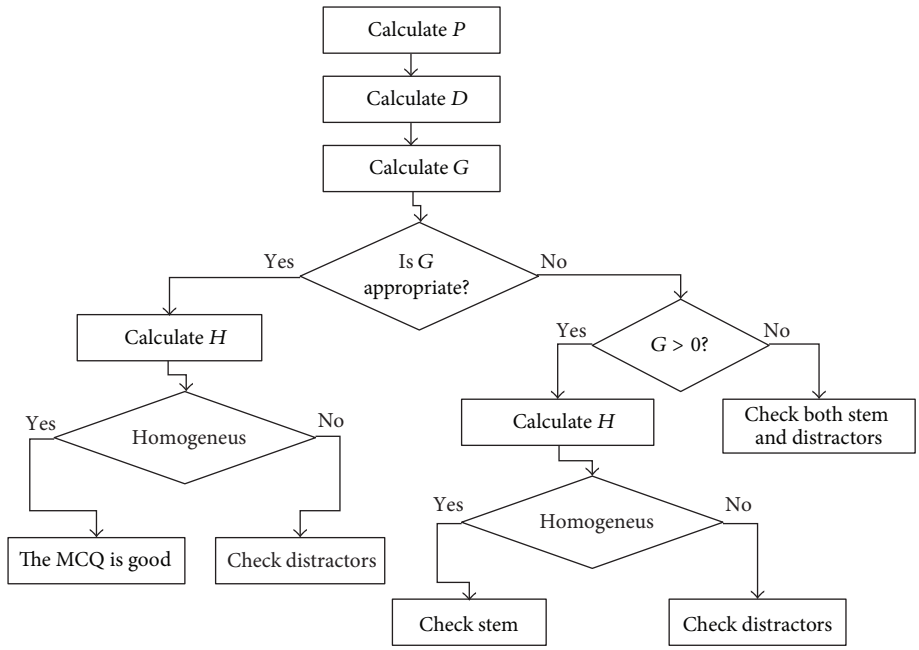

Figure 1: Scheme of the algorithm.

(iii) Items P45 and P46 show two questions with equal difficulty and homogeneity index and different discrimination capacity.

(iv) Items P83 and P123 have very different discrimination index while items P123 and P389 have similar discrimination power with quite different homogeneity levels, perhaps caused by the difficulty of the item. Moreover, the number of errors of item P389 is too high with respect to the number of no answers.

(v) Items P257 and P259 with very different homogeneity index cannot be considered in the same category.

(vi) Finally item P404 is categorized as difficult, but it is good from the point of view of homogeneity and discrimination. The key point for this question is the large number of nonrespondents, perhaps motivated by the interaction of difficulty and homogeneity of the distractors.

All this shows how useful is the combined use of the four indices in order to improve the evaluation process.

Table 5 summarizes the indices classified in the categories specified in Table 3. Most of the questions had a moderate difficulty. Only 7% of the questions were very difficult. For the homogeneity index the approximation to the chi-squared distribution is good when the number of errors is greater than $5(k-1)$, 20 for $k = 5$ (see Table 1 for small values of the number of errors). Almost half of the items in the study had very small number of errors, very much related to easy items. Thus, for 50 students, an item was classified as easy if there were 38 or more correct answers. This means that at most 12 students were distributed among errors and no answers. Anyway, very easy items do not need further analysis if there is no interest in modifying this feature. The quality of the items is quite satisfactory; for instance, more than 75% of the items show a high degree of discrimination. The algorithm provides a categorization oriented to detect the items to be revised. In general 196 of 500 (39.2%) items should be revised and 304 could pass to the database for a posterior use.

Table 5: Results of the MCQs classified in categories.

Index	Classification	Number of items	% over 500 MCQ
Difficulty (P)	Easy	202	40.4%
	Moderate	262	52.4%
	Difficult	36	7.22%
Homogeneity (H)	Homogeneous	355	71%
	Nonhomogeneous	145	29%
Discrimination (D)	High	388	77.6%
	Moderate	51	10.2%
	Low	61	12.2%
No answer (G)	Without E and NA	6	1%
	Errors	44	8.8%
	Good	406	81.2%
	NA	45	9%

Moderate difficulty with high degree of discrimination is the most frequent case in our database.

Finally, the algorithm developed in this work has been implemented in a web application, which can be tested in the web address http://www.med-cr.uclm.es/APEM/index.html.

4. Conclusion

The MCQs are a very common evaluation system in general. In Spain it is used to rank medical students in order to

Table 6: Possible correction for guessing, proportional to the group answers.

	True alternative	Distractor 1	Distractor 2	Distractor 3	Distractor 4	Sum
Group answers	25	3	12	8	2	50
Proportional penalty	0	$K/3$ (8/25)	$K/12$ (2/25)	$K/8$ (3/25)	$K/2$ (12/25)	$25K/24$ (1)

choose specialty. This paper wants to provide a practical tool to achieve high quality in this type of tests used in most of the Spanish schools of medicine and in higher education in general. Thus, it is important to have suitable tools to improve the items and the tests in order to come with a refined database with classified items based on the quality of the distractors.

The results obtained in this work show the utility of the new indices as well as the algorithm developed to detect items and responses to be revised. All this provides a practical tool to create new good items with the requisites wanted by the professor. This allows developing an automated generating system of tests with specific degrees of difficulty and discrimination from a large enough database. Of course this process has to be used as a reference, always supervised by the professor. The main advantages of this are to avoid the risk of building too easy or too difficult tests as well as tests with too low discrimination power. This procedure can help in subjectivity elimination, but it may not be 100%.

The levels used as references for the difficulty and discrimination coefficients come from the literature as well as the performance of the students in the sample. For other situations these levels have to be tuned properly.

Summarizing the results we may say that the new indices in combination with the difficulty index provide a tool to detect inappropriate distractors and sometimes inappropriate items. The discriminating index provides a tool to discard or revise low discriminating items. After this filtering process, an item enters the general database with the corresponding values of the indices. When a new examination has to be run, an automated-supervised process generates suitable tests for the occasion.

There is an old debate about using some kind of penalty for guessing in tests based on MCQs. The usual correction is based upon the assumptions that all wrong answers are guessed wrong and that all correct answers are obtained either by knowledge or guessing. Diamond and Evans [13] offer a thorough review of the topic, stressing advantages and disadvantages. One of the earliest studies in this area was made by Ruch and Stoddard [14] and Ruch and Degraff [15] from different perspectives. Recently Espinosa and Gardeazabal [16] contributed to this discussion with a formal analysis of the effects of penalties. If partial knowledge is taken into account, a penalty based on the general results of the test may be fairer. An example will serve to show this proposal, which needs further consideration, and it is not the aim of this paper. Assume all the students are compelled to answer all the questions. Therefore they will select an alternative at random from the set of alternatives that appear as possibly true for them. Suppose that in a particular question the distribution of answers is shown in Table 6. The optimal choice of the constant K needs a careful study. One is tempted to choose it in such a way that the mean of the four penalties is 1/4, the traditional correction. For example, $K = 24/25$ and the actual corrections appear in Table 6 between parentheses. In this case many examinees selected Distractor 2. The reason may be that either it acts as a good distractor or it is an ambiguous alternative or the students were not well taught in this aspect. In any of these cases the penalty should be minimized. The opposite happens with Distractor 4. The advantage of this method is in considering the whole process of learning and assessing the actual work and knowledge of the examinees. One of the disadvantages may be the possibility of implementing a group strategy, something like selecting always the last question in the set of doubts, but it is unlikely to happen. This is just a proposal to show we are aware of the limitations and advantages of the correction by guessing and as a matter of fact we are working on it. As mentioned above, the procedure of this paper works as a security system where the alarm sounds when there is some probability of some error in the formulation of a particular MCQ. False alarms are not a problem and the number of false alarms can be minimized adjusting conveniently the limit of this probability.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] J. Cangelosi, *Designing Tests for Evaluating Student Achievement*, Addison Wellesley, New York, NY, USA, 1990.

[2] S. M. Case and D. B. Swanson, *Constructing Written Test Questions for the Basic and Clinical Sciences*, National Board of Medical Examiners, Philadelphia, Pa, USA, 3rd edition, 2001.

[3] J. Palés-Argullós, "Cómo elaborar correctamente preguntas de elección múltiple?" *Educación Médica*, vol. 13, no. 3, pp. 149–155, 2010.

[4] J. K. Farley, "The multiple-choice test: writing the questions," *Nurse Educator*, vol. 14, no. 6, pp. 10–12, 1989.

[5] J. Kehoe, "Writing multiple-choice test items," *Practical Assessment, Research and Evaluation*, vol. 4, no. 9, 1995.

[6] M. Tarrant, J. Ware, and A. M. Mohammed, "An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis," *BMC Medical Education*, vol. 9, no. 1, article 40, 2009.

[7] S. Shahbazova and O. Kosheleva, "'Fuzzy' multiple-choice quizzes and how to grade them," *Journal of Uncertain Systems*, vol. 8, no. 3, pp. 216–221, 2014.

[8] A. Oosterhof, "Similarity of various item discrimination indices," *Journal of Educational Measurement*, vol. 13, no. 2, pp. 145–150, 1976.

[9] S.-M. Sim and R. I. Rasiah, "Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper," *Annals of the Academy of Medicine Singapore*, vol. 35, no. 2, pp. 67–71, 2006.

[10] T. L. Kelley, "The selection of upper and lower groups for the validation of test items," *Journal of Educational Psychology*, vol. 30, no. 1, pp. 17–24, 1939.

[11] L. A. Tristrán, "Model for computer-aided item analysis," in *Foro Nacional de Evaluación Educativa*, pp. 45–68, CENEVAL, 1995.

[12] M. H. DeGroot, *Probability and Statistics*, Addison-Wesley, Reading, Mass, USA, 1986.

[13] J. Diamond and W. Evans, "The correction for guessing," *Review of Educational Research*, vol. 43, pp. 181–191, 1973.

[14] G. M. Ruch and G. D. Stoddard, "Comparative reliabilities of five types of objective examinations," *Journal of Educational Psychology*, vol. 16, no. 2, pp. 89–103, 1925.

[15] G. M. Ruch and M. H. Degraff, "Corrections for chance and "guess" vs. "do not guess" instructions in multiple response tests," *Journal of Educational Psychology*, vol. 17, no. 6, pp. 368–375, 1926.

[16] M. P. Espinosa and J. Gardeazabal, "Optimal correction for guessing in multiple-choice tests," *Journal of Mathematical Psychology*, vol. 54, no. 5, pp. 415–425, 2010.