mathematics

Article

Answering Multiple-Choice Questions in Which Examinees Doubt What the True Answer Is among Different Options

Fernando Sánchez Lasheras [1,2,*], José Curbelo [3], Jaime Baladrón Romero [4], Alberto García Guerrero [4], Carmen Peñalver San Cristóbal [4], Tomás Villacampa [5] and Paula Jiménez Fonseca [6]

[1] Faculty of Sciences, University of Oviedo, 33007 Oviedo, Spain
[2] Institute of Space Sciences and Technologies of Asturias (ICTEA), University of Oviedo, 33004 Oviedo, Spain
[3] Faculty of Medicine, University Francisco de Vitoria, 28223 Pozuelo de Alarcón, Spain
[4] Cursos Intensivos MIR Asturias, c/Quintana, 11A, 1, 33009 Oviedo, Spain
[5] Clínica Oftalmológica Villacampa, c/La Cámara, 15, 33401 Avilés, Spain
[6] Medical Oncology Service, Hospital Universitario Central de Asturias, Avenida Roma, 33011 Oviedo, Spain
* Correspondence: sanchezfernando@uniovi.es; Tel.: +34-985-103-388

Abstract: This research explores the results that an examinee would obtain if taking a multiple-choice questions test in which they have doubts as to what the true answer is among different options. This problem is analyzed by making use of combinatorics and analytical and sampling methodologies. The Spanish exam through which doctors become medical specialists has been employed as an example. Although it is difficult to imagine that there are candidates who respond randomly to all the questions of such an exam, it is common that they may doubt over what the correct answer is in some questions. The exam consists of a total of 210 multiple-choice questions with 4 answer options. The cut-off mark is calculated as one-third of the average of the 10 best marks in the exam. According to the results obtained, it is possible to affirm that in the case of doubting over two or three of the four possible answers in certain group questions, answering all of them will in most cases lead to obtaining a positive result. Moreover, in the case of doubting between two answer options in all the questions of the MIR test, it would be possible to exceed the cut-off mark.

Keywords: decision making; multiple-choice test; combinatorics; simulation; education; test performance; random answer; entrance examination

MSC: 05-04; 60-04

Citation: Sánchez Lasheras, F.; Curbelo, J.; Baladrón Romero, J.; García Guerrero, A.; Peñalver San Cristóbal, C.; Villacampa, T.; Jiménez Fonseca, P. Answering Multiple-Choice Questions in Which Examinees Doubt What the True Answer Is among Different Options. *Mathematics* **2022**, *10*, 4543. https://doi.org/10.3390/math10234543

Academic Editors: Silvia Bacci, Fabrizia Mealli and Leonardo Boncinelli

Received: 12 October 2022
Accepted: 29 November 2022
Published: 1 December 2022

1. Introduction

Multiple-choice tests date back at least to the time of the Han dynasty (202 BC–220 AD) in China, when they began to be used in order to select officials for the administration of the empire [1]. In the 20th century, research for the construction and validation of tests started with the works of Binet and Simon [2]. In this century, the Classical Tests Theory [3] and the Item Response Theory [4] were also developed, both becoming popular in many different applications. However, one concern about multiple-choice tests is that they routinely expose students to wrong answers [5]. In four alternative multiple-choice tests, for example, as is the case of the one studied in this paper, three alternatives are wrong, and only one is correct. As examinees read all choices carefully, they read three plausible wrong answers and only one correct answer. Even if subjects pick the correct answer, reading the wrong statements may make those answers seem true later; that is, simply repeating statements increases the probability that those statements will be judged true later [6].

In Spain, all Medicine graduates who want to apply for a position as a specialist doctor must pass the MIR exam [7]. Since 1978, this test has been convened annually, by the Ministries of Health and Education, and it is carried out on the same day and time in various venues, located in different cities in Spain. The date of the exam is published in the

Official State Gazette a few months before it is held. Currently, the test consists of a total of 210 multiple-choice questions, of which 200 correspond to the exam itself, and another 10 are in reserve. These reserve questions will be used to replace those questions that are canceled in the test for any reason.

Each of the questions proposed in the MIR test consists of four possible answer options, of which only one is correct. In order to prevent candidates from choosing to answer questions at random in the hope of getting it right, questions answered incorrectly are penalized with one point, while each correct question contributes three points to the final test grade. That answering incorrectly means a penalty of a third of what corresponds to the correct answer is not accidental, but it is proposed in such a way that, given that each question presents four answer options, a candidate who answered randomly to all the questions of the test should obtain a total score of zero [8,9]. Please also note that in this test, unanswered questions do not score. Thus, the final grade for the test is made up of the product of the correct answers multiplied by three, minus the number of incorrect answers. The total maximum score is therefore 600, and the candidates are classified according to their results in the test. Additionally, in recent exams, a cut-off point is applied whereby those whose grade falls below that cut are excluded. The cut-off mark is calculated as the third of the average of the 10 best marks in the exam, and therefore its value is usually less than 200. For future exams, changes have been announced in its calculation, and it will be 30% of the average of the 10% of the best exams. In practice, a value of about 150 may be assumed.

Although it is difficult to imagine that there are candidates who respond randomly to all the questions of a test exam such as the MIR test, it is common that many of them may doubt what the correct answer is in some questions between two of the four answers or among three answer options or even among all four. In these circumstances, it is reasonable to consider whether the same reward is obtained by leaving doubtful questions blank (neutral effect on the final score) or answering them (adding or subtracting from the final score depending on the correct answer).

Experiments are known [10] in which it has been shown that when in a test with a penalty individuals answer those questions that they consider doubtful, they obtain more points than they lose due to penalties. It has also been verified empirically that in a test, reducing the number of blank questions increases students' overall marks [11]. In the last MIR call, 3975 individuals (33.6% of the total) left 6 or more questions unanswered, and 1429 (12.1%) left more than 20 questions unanswered. This article has been written with the purpose of analyzing what happens when those doctors face a set of questions in which they doubt between two, three, or four answer options to answer those multiple-choice questions at random, as it will have a certain effect on the final points scored in the test. The results are calculated by making use of three different methodologies: analytical functions, combinatorics, and sampling. The three approaches are then compared. Please note that in a real multiple-choice test such as the MIR exam, the number of questions in which examinees have doubts about the right answer should be only a small part of the total.

Although the approach and methodology proposed in this paper can be extended to any test with the total number of questions desired and as many response options as desired, as long as only one of the possible response options is correct, and there is a penalty for wrong answers or unanswered questions, all the results presented are adapted to the characteristics of the MIR exam in 2021, that is, a test with 200 questions and another 10 in reserve, which could replace any of the previous ones in case of cancellation and in which each question consists of 4 answer options of which only one is correct.

This article is structured as follows. After the introduction, which puts into context the usefulness of tests in selection processes, as well as the development of theories dedicated to their study, this paper presents a section on the materials and methods. This section, the longest in the work, describes from a theoretical point of view the penalty necessary to correct the possible effects of chance in a test trial. Next, the expected gain in points is presented, as well as the possible results that would be obtained when answering randomly, hesitating between two, three, or four answer options. Note that although it is already known from previous works that the expected gain in points when hesitating between four options will be zero, this case has also been introduced in order to show that the reasoning followed is similar regardless of the number of answer options among which there is doubt. In addition, in the Material and Methods section (Section 2), the characteristics of the server used to carry out the calculations are also specified, as well as the software in which the numerical simulations have been programmed. Next, the Results and Discussion section (Section 3) is presented, in which the results of the numerical simulations carried out are described and analyzed; following this, a brief section is included with the main conclusions of the study, and, finally, there is an annex with the source code used to carry out the aforementioned numerical simulations.

2. Materials and Methods

2.1. Theoretical Background

To assess a future medical specialist's performance, one must evaluate their medical knowledge. This rule can be applied to any other professional career. An effective way to assess a learner's comprehension of a particular subject is to use a well-crafted multiple-choice questionnaire. In recent years, some research providing guidelines for creating high-quality multiple-choice questions has been published [12,13]. These works even provide methodologies able to produce multiple-choice questionnaires automatically [13] that can later be fine-tuned by human experts in order to ensure proper syntax, meaning, and clarity.

According to previous research, one of the most outstanding works to take into account in order to create high-quality multiple-choice questionnaires is the one published in 1989 by Haladyna and Downing [14]. It is true to say, however, that at this moment in time, despite the large number of interesting papers [15,16] that are focused on how to create high-quality questions for multiple-choice questionnaires, there is a lack of statistical analysis of how individuals perform when answering multiple-choice tests.

The studies currently available are in most cases focused on the analysis of the quality of the single best answer of multiple-choice questions used in different fields such as pharmaceutics exams [17]. In the case of this research, its aim was to calculate for this purpose the difficulty index, discriminating index, and distractor efficiency and analyze the relationship between them and the number of answer options to multiple-choice questions and their cognitive level. Other works, such as the paper by Kumar et al. (2021) [18], focused on checking some questions on medical students, assessing their performance, and identifying good and ideal multiple-choice questions which could be part of the question bank for future use in medical tests.

As far as the authors know, and after having performed a literature review, it can be said that there is no other article that works on the same topic as the one presented here. Nevertheless, there are some works that may be considered of interest and that are in a certain way related to the present research. One of the most remarkable of these is the work published by Burton in 2001 [19], which deals with the topic of quantifying the effects of chance in multiple-choice and true-or-false tests. This research focuses on describing the weakness of this kind of test and proposes four measures of test unreliability that quantify the effects of question selection and guessing, both separately and together.

Finally, another empirical study [20] that considers the effect of the order of questions concluded that that were multiple-choice questions ordered according to their difficulty from easiest to most difficult or vice versa, the order would not affect the test performances of the examinees, no matter what their level of knowledge was. Please also note that this kind of study that makes use of simulations to perform optimizations is common in many fields of science and technology [21–23].

2.2. Necessary Penalty for the Correction of Chance in a Test

In this section, the necessary penalty is deducted so that in a test in which each item has m answers, only one of which is correct, answering the questions randomly results in a total average score of zero points. Please note that all the equations that are deduced in this section and in those following only required basic knowledge of statistics. Thus, if each item presents m response alternatives of which only one is correct, the probability of hitting it by chance would be one in m:

$$P(A) = \frac{1}{m} \tag{1}$$

That is, as is expressed in Equation (1), they will be right 1 out of every m times that they are answered randomly. Therefore, the probability of making a mistake in answering is given by the formula presented in Equation (2):

$$P(E) = \frac{m-1}{m} \tag{2}$$

If this is applied to a set of n questions, $\frac{n}{m}$ of them will be correct, and $n\cdot\frac{m-1}{m}$ wrong. If each correct question provides k points and the incorrect question penalizes p points, the sum of points obtained by answering all those questions randomly is equal to:

$$S = k\cdot n\cdot\frac{1}{m} + p\cdot n\cdot\frac{m-1}{m} \tag{3}$$

As the score obtained when Equation (3) is applied should be zero, the formula is as follows:

$$0 = k\cdot n\cdot\frac{1}{m} + p\cdot n\cdot\frac{m-1}{m} \tag{4}$$

$$\begin{gathered}\text{Taking into account that } m \neq 0: \\ 0 = k\cdot n + p\cdot n\cdot(m-1)\end{gathered} \tag{5}$$

$$p = -\frac{k}{m-1} \tag{6}$$

When Equation (4) is operated in order to obtain the penalty value, Equation (5) is obtained. Penalization p for each failed question is presented in Equation (6). In the case of the present research, where $m = 4$, the penalization should be $-\frac{k}{3}$ for each question incorrectly answered. Please note that as in our case, each question answered correctly provides $k = 3$ points, the penalty should be -1.

2.3. Expected Score When Answering Randomly, Hesitating between Two of the m Possible Answers of a Test Question

Let n be the number of questions of a test with m answer options in which there is doubt between two of these options. If the number of questions is large enough, and an individual answers them randomly, it can be assumed that the number of correct answers will be of $\frac{n}{2}$ and that the number of mistakes will be also of $\frac{n}{2}$. Therefore, given that k points are awarded for each correct answer, the score that will be obtained for the questions answered correctly will be $k\cdot\frac{n}{2}$, while the penalty that will be applied for answering $\frac{n}{2}$ questions incorrectly will be of $p\cdot\frac{n}{2}$. In other words, the balance of points obtained for answering these n questions will be $\frac{n}{2}(k+p)$.

Therefore, any individual that answers a total of n questions, each of which has m answer options and with each one of these questions providing k points when answered correctly and each question answered incorrectly penalizing according to the result obtained in Equation (6), will obtain a total number of points given by $\frac{n}{2}\left(k - \frac{k}{m-1}\right)$. When this formula is applied to the case of the present research, in which each question has $m = 4$ different answer options and considering that the number of points given by each question correctly answered is $k = 3$ and the number of points given by each question answered incorrectly is $p = -1$, the total amount of points obtained would be n.

2.4. Possible Results When Doubting between Two of the Four Possible Answers of a Test Question

In the particular case of the problem under study, given a test question with four possible answer options in which there is doubt between two of them, if one of the two answers is chosen randomly, the probability of choosing the right one and, therefore, of obtaining 3 points is 50%, while the probability of failing and getting a -1 penalty will also be 50%. In the same way, if there are 2 questions in which there is doubt between 2 possible answers, calling c the correct answer and i the incorrect answer, if answered randomly, there are 4 possible variations in the result, as is shown in Table 1.

Table 1. Possible results that would be obtained when answering two questions in a test in which there is a doubt between two options over which is the correct answer.

Question 1	Question 2	Points
c	c	6
c	i	2
i	c	2
i	i	−2

Therefore, Table 1 shows the variations with the repetition of 2 elements taken 2 by 2, which is the statistical problem that results from doubting between 2 possible answer options in 2 test questions. As can be seen in Table 1, there is one result among the 4 possible in which a negative score is obtained, while in the other three cases, the score obtained is positive. Please also note that avoiding the risk, that is, the alternative of leaving the questions unanswered, would return a result of zero.

In the same way, if there is doubt between two possible answers in three questions, a scenario with 8 answer possibilities is presented. Please see in Table 2 how of these 8 possible outcomes, in only 1 (12.5%) is a negative score obtained. Therefore, as can be observed in the table referred to, in one of the eight possible cases 9 points are obtained, while there are three cases in which five points are obtained and another three in which only one point is obtained.

Table 2. Possible results obtained when answering three questions in a test in which there is a doubt between two options.

Question 1	Question 2	Question 3	Points
c	c	c	9
c	c	i	5
c	i	c	5
c	i	i	1
i	c	c	5
i	c	i	1
i	i	c	1
i	i	i	−3

This way of working can be extended to any number of questions, and, in the same way, Table 3 shows the results obtained for the case in which there is doubt between two options in four questions. As can be seen in the aforementioned table, in only one of the 16 possible cases (6.25%) would the score obtained be negative, although there would be four other cases (25%) in which the score would be equal to 0.

Table 3. Possible results obtained when answering four questions in a test in which there is a doubt between two options over which the right one is.

Question 1	Question 2	Question 3	Question 4	Points
c	c	c	c	12
c	c	c	i	8
c	c	i	c	8
c	c	i	i	4
c	i	c	c	8
c	i	c	i	4
c	i	i	c	4
c	i	i	i	0
i	c	c	c	8
i	c	c	i	4
i	c	i	c	4
i	c	i	i	0
i	i	c	c	4
i	i	c	i	0
i	i	i	c	0
i	i	i	i	−4

Table 4 shows the results obtained for the case of 5 questions. Here, a total of 6 out of 32 cases (18.75%) have a negative score. Note that since the number of questions is odd, a score of zero is never reached. In addition, it should also be noted that although the trend is that as the number of questions in which there is doubt between two answers increases, the percentage of cases in which a negative score is obtained decreases. This decrease is not linear and is affected by the number of questions being considered. In other words, when we go from considering n questions to considering $n + 1$, the percentage of cases in which a negative score would be obtained does not decrease for all values of n, but rather, if n is a multiple of 4, the percentage of cases in which negative scores are obtained increases by $n + 1$.

Thus, given a test with four possible answer options and in which 3 points are awarded for each correct answer, and each incorrect answer is penalized with -1 or with any other quantity wherein the penalty is equal to one-third of the score awarded for each correct answer, and being $n \in \mathbb{N}$ the number of questions of said test in which there is doubt between two possible answers, both being considered equally likely, the total number of possible different ways of answering said questions is 2^n. Of these possibilities, as has been verified through the previous examples, there will be some in which the score balance obtained will be negative. This situation will occur when the number of questions answered incorrectly is greater than three times the number of questions answered correctly. Thus, for example, in the case of 12 questions, negative scores will be obtained when 10, 11, or 12 of the answers to the questions are incorrect. Likewise, the balance will be zero points when nine answers are incorrect and three correct; in the rest of the cases, the score will be positive.

For the proposed case of 12 questions, a negative result will be obtained in all possible permutations with repetition of two elements (incorrect answers and correct answers) in which one of them, the incorrect answer, is repeated 10 times, 11 times, or 12 times and the other, the correct answers, are repeated two, one or zero times. Therefore, the generalized formula that determines the number of cases in which a negative score is obtained is given by:

$$\sum_{i=1}^{k} \frac{n!}{(n-i)! \cdot i!} + 1, \text{ where } k = truncate\left(\frac{n-1}{4}\right) \tag{7}$$

Table 4. Possible results obtained when answering five questions in a test in which there is a doubt between two options over which the right one is.

Question 1	Question 2	Question 3	Question 4	Question 5	Points
c	c	c	c	c	15
c	c	c	c	i	11
c	c	c	i	c	11
c	c	c	i	i	7
c	c	i	c	c	11
c	c	i	c	i	7
c	c	i	i	c	7
c	c	i	i	i	3
c	i	c	c	c	11
c	i	c	c	i	7
c	i	c	i	c	7
c	i	c	i	i	3
c	i	i	c	c	7
c	i	i	c	i	3
c	i	i	i	c	3
c	i	i	i	i	−1
i	c	c	c	c	11
i	c	c	c	i	7
i	c	c	i	c	7
i	c	c	i	i	3
i	c	i	c	c	7
i	c	i	c	i	3
i	c	i	i	c	3
i	c	i	i	i	−1
i	i	c	c	c	7
i	i	c	c	i	3
i	i	c	i	c	3
i	i	c	i	i	−1
i	i	i	c	c	3
i	i	i	c	i	−1
i	i	i	i	c	−1
i	i	i	i	i	−5

Equation (7) is valid for any n greater than or equal to 5. In addition, if the number of questions in which there is doubt between two answers is a multiple of 4, there are some cases in which the total score obtained would be equal to zero. The number of possibilities in which this occurs is given by the formula:

$$\frac{n!}{\left(\frac{n}{4}\right)! \cdot \left(\frac{3 \cdot n}{4}\right)!} \tag{8}$$

Formula (8) is obtained by taking into account that these are permutations with repetition in which the incorrect answers are present in $\frac{3 \cdot n}{4}$ of the total number of questions and the correct ones in $\frac{n}{4}$ and since n is a multiple of four, the results of both operations are integers. Note also that in the case where the number of questions n is not a multiple of four, it is not possible to achieve a score of zero, and, therefore, Formula (8) is not applicable.

2.5. Expected Score When Answering Randomly, Hesitating between Three of the Possible Answers of a Test Question

Let n be the number of questions in a test with m answer options in which there is doubt over what the right answer is among three options. If the number of questions is large enough, it can be assumed that the number of correct answers will be $\frac{n}{3}$, while the number of failed questions will be $n - \frac{n}{3}$. Therefore, if k points are given for each question answered correctly, the score that will be obtained for the questions answered correctly will be $k\frac{n}{3}$, while the penalty that will be applied for answering incorrectly $n - \frac{n}{3}$ questions will be of $\frac{-k}{m-1}\left(n-\frac{n}{3}\right)$. Therefore, the balance of points obtained for answering these n questions will be of $k\frac{n}{3}-\frac{2}{3}\frac{kn}{m-1}$. In the case under study, with $m=4$ and $k=3$ $\frac{n}{3}$ points are obtained, and taking into account that the maximum score attainable with n questions is $3n$, on average, one-ninth of the total score will be obtained.

2.6. Possible Results When Doubting between Three of the Four Possible Answers of a Test Question

Considering a test question in which three of the four possible answers are in doubt and to which a penalty of -1 is applied for each wrong answer and 3 points are awarded for each correct answer, if the correct answer is designated as c and the two incorrect answers as i_1 and i_2, then the possibility of obtaining a positive score as a result of answering that question is 33.33%.

Working in a similar way to what was done in the section relating to the doubts between 2 answers, Table 5 presents the results obtained for two questions when there is doubt among three options. These are the variations with repetition of 3 elements taken 2 by 2, and, therefore, a total of 9 different possible results are presented. Note that an equivalent approach to this problem would be to consider that there is a correct answer c and a single incorrect answer i and that, while the correct answer is chosen with a probability of $\frac{1}{3}$, the wrong answer is chosen with a probability of $\frac{2}{3}$. In view of the results shown in Table 5, in 5 of the 9 cases (55.55%), the score achieved would be positive, while in the remaining 4 (44.44%), it would be negative.

Table 5. Possible results that would be obtained when answering two questions in a test in which there is a doubt between three options over which is the correct answer.

Question 1	Question 2	Points
c	c	6
c	i_1	2
c	i_2	2
i_1	c	2
i_1	i_1	-2
i_1	i_2	-2
i_2	c	2
i_2	i_1	-2
i_2	i_2	-2

Similarly, Table 6 shows the possible results that would be achieved for the case of three test questions in which there is doubt in all of them, in an equiprobable way between three of the four answers available in each of them. Thus, in this case, of the 27 possible cases, a negative score would be obtained in 8 of them (29.63%) and a positive score in 19 (70.37%), and it is not possible to achieve a score equal to zero.

Generalizing what has been stated in the previous cases, for a total of n questions in which there is doubt among three answer options over which is correct, the total number of possible different answers to the set of said questions is equal to 3^n. Of these answers, if you want to know which ones result in a negative score, it is possible to use Formula (7). When making use of Formula (7), it must be taken into account that there are two incorrect answer options, called i_1 and i_2. If the same nomenclature is used to refer to the number of times each of the wrong answers is chosen, the above formula must be modified by replacing $i!$ for all the possible products of $i_1!\cdot i_2!$ such that they verify that $i_1+i_2=i$. In addition, the case that was previously unique, that for all the questions the answer chosen to be wrong, now becomes 2^n possibilities, since each incorrect question can be due to either choosing the wrong answer i_1, or i_2.

Table 6. Possible results that would be obtained when answering three questions in a test in which there is doubt between three options over which is the correct answer.

Question 1	Question 2	Question 3	Points
c	c	c	9
c	c	i_1	5
c	c	i_2	5
c	i_1	c	5
c	i_1	i_1	1
c	i_1	i_2	1
c	i_2	c	5
c	i_2	i_1	1
c	i_2	i_2	1
i_1	c	c	5
i_1	c	i_1	1
i_1	c	i_2	1
i_1	i_1	c	1
i_1	i_1	i_1	−3
i_1	i_1	i_2	−3
i_1	i_2	c	1
i_1	i_2	i_1	−3
i_1	i_2	i_2	−3
i_2	c	c	5
i_2	c	i_1	1
i_2	c	i_2	1
i_2	i_1	c	1
i_2	i_1	i_1	−3
i_2	i_1	i_2	−3
i_2	i_2	c	1
i_2	i_2	i_1	−3
i_2	i_2	i_2	−3

For example, in the case of 5 questions, negative scores would only be obtained if 4 or 5 of them were answered incorrectly. Therefore, the number of possibilities that would correspond to negative scores would be given by Equation (9):

$$\begin{aligned}
&\frac{5!}{(5-4)!0!4!}+\frac{5!}{(5-4)!1!3!}+\frac{5!}{(5-4)!2!2!}+\frac{5!}{(5-4)!3!1!}+\frac{5!}{(5-4)!4!0!}+2^5\\
&=\frac{5\cdot4\cdot3\cdot2\cdot1}{1!0!4\cdot3\cdot2}+\frac{5\cdot4\cdot3\cdot2\cdot1}{1!1!3\cdot2\cdot1}+\frac{5\cdot4\cdot3\cdot2\cdot1}{1\ 2\cdot1\ 2\cdot1}+\frac{5\cdot4\cdot3\cdot2\cdot1}{1!\ 3\cdot2\cdot1\cdot1}\\
&+\frac{5\cdot4\cdot3\cdot2\cdot1}{1!\ 4\cdot3\cdot2\cdot1\cdot1\cdot0!}+2^5=5+50+30+20+5+2^5=80+32\\
&=112
\end{aligned} \tag{9}$$

If, for example, there are 12 questions in which there is doubt over three answer options, in those cases in which the sum of i_1 and i_2 is greater than 9, that is, 10, 11, or 12, the score obtained will be negative, and the total number of cases in which this assumption occurs out of the possible total of $3^{12} = 531,411$ will be equal to 96,256.

In the same way as what happened when doubting which is the right choice between two possibilities, it is only possible to achieve results equal to zero when the number of questions is a multiple of four; that is, a score of zero is only reached when a quarter of the questions are correct and the other three-quarters are answered incorrectly. However, in this case, the mistake can be caused by choosing either the wrong answer i_1 or i_2. Therefore, given that in each of the incorrect answers there are two options, the number of possibilities of answering and obtaining a result equal to zero for n questions, n being a multiple of 4, is equal to the formula determined for two answer $2^{3\cdot\frac{n}{4}}$ options but multiplied by $2^{3\cdot\frac{n}{4}}$. This formula is presented in Equation (10):

$$2^{3\cdot\frac{n}{4}}\cdot\frac{n!}{\left(\frac{n}{4}\right)!\cdot\left(\frac{3\cdot n}{4}\right)!} \tag{10}$$

2.7. Expected Score When Answering Randomly, Doubting over All the Possible Answers of a Test Question

Let n be the number of questions of a test with m answer options where all of them are considered with the same probability of being the correct one. It can be assumed that the number of correct answers will be $\frac{n}{m}$, while the number of incorrect questions will be $n-\frac{n}{m}$. Therefore, if k is the number of points given by each correct answer, the score that will be obtained for the questions answered correctly will be $\frac{kn}{m}$, while the penalty that will be applied for answering incorrectly $n-\frac{n}{m}$ questions will be $\frac{k}{m-1}\cdot\left(n-\frac{n}{m}\right)$. Therefore, the balance of points obtained by answering these n questions will be equal to zero. Given that the penalty applied is exactly what is necessary for the random answer to all the questions to yield a result equal to zero, this result is the expected one. Therefore, there is no need to make an analysis of this case.

Please note that the expected score that will be obtained when answering randomly with doubts over all four of the four possible answers to a test question does not depend on the sample size, that is, the number of examinees that take the test. The same occurs with all the other cases that are described in the Materials and Methods section from Sections 2.2–2.6.

2.8. Hardware and Software

All the calculations of this work were performed with a computer with the Ubuntu 18.04.5 LTS operating system installed. This machine is equipped with an Intel® Core™ i7-7700K CPU @ 4.20 GHz and 64 GB of RAM. For the programming of the source code, the statistical software R was used in version 3.6.3 [24].

3. Results and Discussion

This section presents some results that are considered to be of interest in the context of a test exam such as MIR, consisting of a large number of questions and in which, in some of them, the examinee may hesitate between two or three of the four possible answer options. Note that since in the Materials and Methods section (Section 2) it has been shown that in the case of doubting over the four answer options for a number of questions n, the average score obtained by answering all of them tends to be zero, this case is not analyzed in this section.

It should be noted that, if the intention is to analyze the results obtained in the MIR test by a student who doubts between 2 answers in all the questions of this test without taking into account the reserve ones, a total of 2^{200} variations with repetition should be considered, which would mean that there could be more than $1.6\cdot10^{60}$ different ways to answer the exam.

Following a similar reasoning, in the case of doubting over all the test questions among 3 options, 3^{200} possibilities should be taken into account, which means more than $2.6\cdot10^{95}$ possible exams with different results. Despite the speed of computers, an exhaustive calculation of all these cases is not possible. Therefore, in this section, it is proposed to obtain an approximation to the results achieved in the MIR test through a sample that analyzes subsets of exams determined randomly for each of the cases that arise.

3.1. Results When Doubting between Two of the Four Possible Answers of a Test Question

Taking into account the approach taken in the Materials and Methods section (Section 2), in order to find the expected results when doubting over a set of n questions between two possible answers of the four that are offered, it is possible to use the analytical formulae deduced in said section. Specifically, from Formula (7) it is possible to calculate the cases with negative scores, from Formula (8) to find the cases with scores equal to zero as long as the number of questions is a multiple of four, and, finally, the number of cases with positive scores can be obtained by subtracting from 2^n the number of cases with negative and zero scores. Table 7 shows the total number of responses with sums of negative, zero, and positive scores and the percentage of negative results over the total for any number of questions between 1 and 23 when in doubt between 2 of the possible response options. Note that the values in this table have been obtained using the analytical formulae mentioned in this paragraph. As a summary of the results shown in Table 7, it can be said that when doubting between two of the possible answers in 14 questions or more, if the examinee answers randomly to all of them, the probability of obtaining a negative score is below 5%.

Table 7. Total responses with negative, zero, and positive scores and percentage of negative scores for any number of questions between 1 and 23 when doubting between two of the possible response options.

Num. Quest.	Answers				
	Negative	Zero	Positive	Tot. Answers	% Negative
1	1	0	1	2	50.00%
2	1	0	3	4	25.00%
3	1	0	7	8	12.50%
4	1	4	11	16	6.25%
5	6	0	26	32	18.75%
6	7	0	57	64	10.94%
7	8	0	120	128	6.25%
8	9	28	219	256	3.52%
9	46	0	466	512	8.98%
10	56	0	968	1024	5.47%
11	67	0	1981	2048	3.27%
12	79	220	3797	4096	1.93%
13	378	0	7814	8192	4.61%
14	470	0	15,914	16,384	2.87%
15	576	0	32,192	32,768	1.76%
16	697	1820	63,019	65,536	1.06%
17	3214	0	127,858	131,072	2.45%
18	4048	0	258,096	262,144	1.54%
19	5036	0	519,252	524,288	0.96%
20	6196	15,504	1,026,876	1,048,576	0.59%
21	27,896	0	2,069,256	2,097,152	1.33%
22	35,443	0	4,158,861	4,194,304	0.85%
23	44,552	0	8,344,056	8,388,607	0.53%

Another way of obtaining the same values, which would serve as a verification of the proposed formulae, consists of calculating all the variations with repetition of two elements, correct answers (c) and incorrect answers (i), taken from n to n for afterwards calculating in which cases the result is positive, null, or negative. Thus, Table A1 of Appendix A shows the source code that calculates for any number of questions between 1 and 200 the total number of cases in which the score obtained is negative, null, and positive. Although this code saves the need to use the analytical Formulae (7) and (8), the time required for the calculation grows considerably as the value of n increases. Given the existence of the aforementioned analytical formulae, it is not essential to use this calculation methodology, but its presentation in this article is of interest, since it offers an alternative form that could be useful in obtaining other results such as, for example, the percentage of cases in which a score greater than 30% of the maximum attainable is obtained, without the deduction of new analytical formulae.

Given that this technique requires a high computation time that increases considerably when the number of questions grows, another way of simulation is also proposed. In this method, a million replications of a test are calculated in which questions have been answered randomly, hesitating between two answers: one incorrect and another correct.

The source code of this simulation can be found in Table A2 of Appendix A, showing the number of cases in which the total score obtained is negative, null, and positive.

Table 8 presents the percentage of responses with a negative score sum calculated by means of the analytical formula, as well as the time required to calculate all the possible combinations. That same result is calculated with the help of a computer (permutation method), in order to later show the percentage of questions with negative scores obtained by the sampling technique when 1 million repetitions are applied, indicating the time required for its calculation. It is observed how from the 22 questions, when the number of permutations is exhaustively calculated, the time needed is multiplied on average by 4, while when applying the sampling methodology, times are multiplied by only 1.05. Thus, the objective of this table is to show three methods of obtaining the same results. These methods are as follows:

- The **analytical formula**, whose main advantage is accuracy, but whose main disadvantage is that if another kind of result is sought after, a new formula should be deduced. For example, what the probability is of obtaining scores higher than a third of the maximum.
- The **use of combinatorics**, which, although it saves the difficulty of needing to deduce new analytical formulae, has as a drawback the non-linear growth of calculation time as the number of questions considered increases and, finally,
- The **use of sampling**, which, although it has the drawback of a lack of accuracy, has the advantages of both its versatility to adapt to the type of problem that arises at any time and the execution time, since this can be controlled through the number of repetitions that are imposed on the experiment. Note that in this research, a million repetitions of each experiment are performed.

Table 8. Percentage of responses with negative sums of scores calculated by means of the analytical formula, time required to calculate all possible combinations by means of the combinatorics method, percentage of negative questions calculated using sampling with 1 million repetitions, and time required for its calculation.

Num. Quest.	Analytic Function	Combinatorics	Sampling	Sampling
	% Negative	Time (s)	% Negative	Time (s)
1	50.00%	0		
2	25.00%	0	25.02%	7289.886
3	12.50%	0	12.45%	10,089.784
4	6.25%	0	6.24%	11,061.708
5	18.75%	0	18.73%	10,562.268
6	10.94%	0	10.98%	12,534.918
7	6.25%	0	6.26%	14,274.719
8	3.52%	0.001	3.53%	18,137.062
9	8.98%	0.002	8.99%	21,261.337
10	5.47%	0.005	5.48%	23,201.416
11	3.27%	0.021	3.26%	27,024.192
12	1.93%	0.054	1.93%	28,489.992
13	4.61%	0.181	4.67%	27,567.958
14	2.87%	0.694	2.84%	29,874.998
15	1.76%	2.702	1.76%	32,038.688
16	1.06%	10.815	1.07%	34,471.178
17	2.45%	45.08	2.44%	35,859.301
18	1.54%	195.963	1.56%	42,500.085
19	0.96%	810.752	0.97%	44,024.053
20	0.59%	3245.89	0.59%	47,458.234
21	1.33%	13,088.809	1.33%	49,053.421
22	0.85%	52,314.594	0.83%	43,743.01
23	0.53%	283,071.635	0.54%	46,038.456

The comparison of the results shown in Table 7 with those obtained in Table 8 serves as a comparison of the equivalence between the methodologies applied.

Table 9 shows the results obtained for the percentage of tests with negative sum scores for a number of questions between 10 and 100 calculated from 10 to 10 and using the sampling methodology with 1 million repetitions for each number of questions. As can be seen in the aforementioned table, when the number of questions is 90, the percentage of cases in which a negative score has been obtained is 0.0001%, that is, one in a million and, therefore, although for 100 questions, there is a chance of getting negative scores, the frequency of such an event is less than one in a million. Note that the alternative strategy for this subset of questions would be to leave them unanswered, and in such cases, the gain would be zero.

Table 9. Percentage of responses with negative sums of scores for a number of questions between 10 and 100 calculated using sampling with 1 million repetitions and the time required for its calculation.

Num. Quest.	Sampling	Sampling
	% Negative	Time (s)
10	5.48%	23,201.416
20	0.59%	47,458.234
30	0.2582%	74,080.827
40	0.0360%	110,463.663
50	0.0157%	132,601.933
60	0.0018%	159,242.671
70	0.0008%	193,324.229
80	0.0002%	210,989.068
90	0.0001%	228,728.687
100	0%	251,057.638

Figure 1 represents the numerical values of Tables 10 and 11 in relation to the percentage of tests in which a negative score would be obtained. As can be seen in this figure, the percentage of cases in which negative scores are reached decreases rapidly as the number of questions increases. It may also be observed that it is not always true that this percentage, although it decreases, is lower for a number of questions $n + 1$ when compared to n, but if n is a multiple of four, for $n + 1$, the percentage of cases in which a negative score would be reached becomes greater than for n. This is so given that when the number of questions is a multiple of four, there is also the possibility that the total score achieved will be zero.

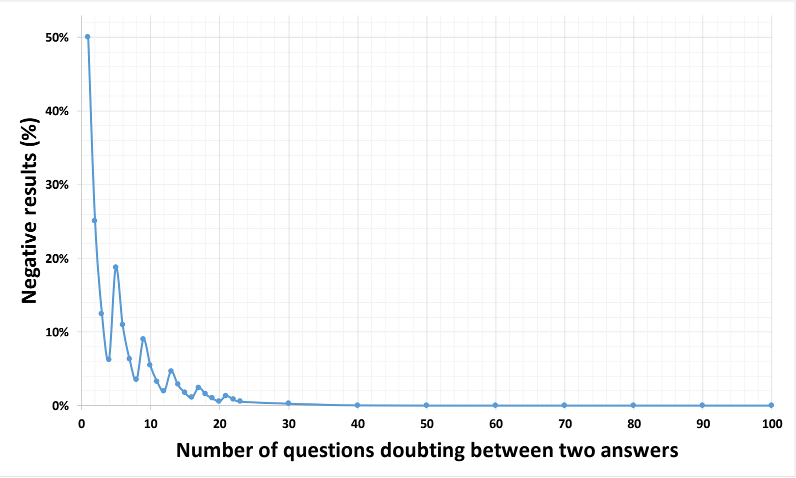

Figure 1. Percentage of negative results when doubting between two answers in a number of questions between 1 and 100.

Table 10. Simulation of the results that would be obtained when carrying out a million tests of 200 questions in which everyone hesitates between two answer options.

Points	N	Accumulated	Accumulated Percentage
70	1	1	0.0001%
80	2	9	0.0009%
90	13	37	0.0037%
100	106	240	0.0240%
110	284	702	0.0702%
120	1034	2836	0.2836%
130	2199	6538	0.6538%
140	5909	19,990	1.9990%
150	10,582	38,459	3.8459%
160	20,719	89,411	8.9411%
170	29,886	144,230	14.4230%
180	43,955	261,989	26.1989%
190	51,950	361,834	36.1834%
200	56,424	528,535	52.8535%
210	54,366	638,808	63.8808%
220	43,671	781,891	78.1891%
230	34,381	855,623	85.5623%
240	20,576	931,322	93.1322%
250	13,453	961,444	96.1444%
260	5912	985,881	98.5881%
270	3089	993,361	99.3361%
280	1075	998,224	99.8224%
290	404	999,314	99.9314%
300	108	999,871	99.9871%
310	33	999,952	99.9952%
320	5	999,994	99.9994%
330	3	1,000,000	100%

Table 11. Total responses with sums of negative, zero, and positive scores and percentage of negative questions for any number of questions between 1 and 15 when in doubt among three of the possible response options.

Num. Quest.	Answers				
	Negative	Zero	Positive	Tot. Answers	% Negative
1	2	0	1	3	66.67%
2	4	0	5	9	44.44%
3	8	0	19	27	29.63%
4	16	32	33	81	19.75%
5	112	0	131	243	46.09%
6	256	0	473	729	35.12%
7	576	0	1611	2187	26.34%
8	1280	1792	3489	6561	19.51%
9	7424	0	12,259	19,683	37.72%
10	17,664	0	41,385	59,049	29.91%
11	41,472	0	135,675	177,147	23.41%
12	96,256	112,640	322,545	531,441	18.11%
13	514,048	0	1,080,275	1,594,323	32.24%
14	1,249,280	0	3,533,689	4,782,969	26.12%
15	3,002,368	0	11,346,539	14,348,907	20.92%

In the case of the MIR test, which consists of 200 questions and another 10 in reserve that may replace any of the original ones if there are any cancellations, it is interesting to consider, for example, what the probability is of exceeding the cut-off mark in the case of hesitating between two answers in all the exam questions or, also, what the probability would be of obtaining 50% of the total points by hesitating between two answers in all the questions of the exam. Note that the cut-off mark is set at one-third of the average score obtained in the test for the ten best exams and is, therefore, somewhat less than a third of the total score that would be obtained by answering all the questions correctly.

In order to answer these questions, we have simulated the scores that would be obtained if a million tests were performed in which, in all the questions, there was doubt between two answer options. The code used for this simulation is shown in Table A3 of Appendix A. Table 10 shows the results obtained by carrying out a million tests of 200 questions each, in which all of them hesitate between two answer options. Therefore, for example, given the information in this table, in 52.8535% of the cases, the score obtained would be equal to or fewer than 200 points, and, therefore, in 47.1465% of the cases, the score obtained would be greater than 200 points, which is a third of the maximum score attainable in a 200-question exam in which three points are awarded for each correct answer. The Spanish Health Ministry recently changed the calculation of the cut-off mark for the 2022 MIR exam. From now on, instead of being calculated on the ten best exams, it will be calculated on 10% of the best exams. In addition, instead of being a third of the average of the best exams, it will be 25%. In practice, this will mean a cut-off mark of about 150, whose probability of being reached or exceeded when doubting between two options in all questions is 96.1541%.

3.2. Results When Doubting between Three of the Four Possible Answers of a Test Question

Taking into account the approach made in the Materials and Methods section (Section 2), to find out the results expected when doubting in a set of n questions among three possible answers out of four, it is possible to make use of the analytical formula. Table 11 shows the total number of responses with sums of negative, zero, and positive scores and the percentage of negative questions over the total for any number of questions between 1 and 15. Note that the values in this table have been obtained by means of the analytical formulae. This means that for example, when an examinee has doubts over three possible response options in 10 different questions, and they answer them randomly, there is a 29.91% chance of obtaining a negative score. In this case, from a probabilistic point of view, there are 59,049 different possible answer combinations to the 10 questions, of which 17,664 would give a negative score, while 41,385 would return a positive score.

When doubting between two response options, another way of obtaining these same values that would serve to verify the proposed formulas would be to obtain all the variations with repetition of three elements taken from n to n to subsequently calculate in which cases results are positive, null, or negative.

Table A4 of Appendix A shows the source code that calculates the total number of cases in which the score obtained is negative, null, and positive for any number of questions between 1 and 200. Table 12 presents the percentage of responses with negative sums of scores calculated by means of the analytical formula, as well as the time required to calculate all the possible combinations and obtain the same values with the help of a computer. It also shows the percentage of questions with negative scores obtained when the sampling technique with 1 million repetitions is applied and indicates the time required for its calculation. It can be seen how, after 14 questions, the time required for sampling with 1 million repetitions is less than that required for the exhaustive calculation of all permutations. The code used to calculate the results using sampling is presented in Table A5 of Appendix A. The interest of this table is to highlight how the application of one methodology or another would be more favorable, depending on the scenario.

Table 12. Percentage of responses with negative results calculated by means of the analytical formula, time required to calculate all possible combinations, percentage of negative questions calculated using sampling with 1 million repetitions, and time required for its calculation.

Num. Quest.	Analytic Function	Combinatorics	Sampling	Sampling
	% Negative	Time (s)	% Negative	Time (s)
1	66.67%	0		
2	44.44%	0	44.39%	11,209.990
3	29.63%	0.001	29.62%	15,337.608
4	19.75%	0.001	19.74%	20,328.257
5	46.69%	0.001	46.07%	15,497.987
6	35.12%	0.005	35.11%	14,870.793
7	26.34%	0.03	26.32%	21,810.212
8	19.51%	0.147	19.54%	26,447.580
9	37.72%	1.063	37.70%	35,653.45
10	29.91%	8.998	29.92%	34,366.795
11	23.41%	97.994	23.37%	36,871.024
12	18.11%	1294.693	18.07%	43,100.296
13	32.24%	11,879.693	32.17%	44,850.212
14	26.12%	11,5925.395	26.06%	44,210.714

Table 13 shows the results obtained relative to the percentage of responses with negative sum scores for a number of questions between 10 and 200, calculated from 10 to 10 and using sampling with 1 million repetitions for each number of questions. Note that after 160 questions, the percentage of cases in which a negative score has been obtained is below 1%. To calculate this table, the sampling methodology has been preferred given the time that would be required if instead of sampling the combinatorics methodology were to be applied.

Table 13. Percentage of responses with negative sums of scores for a number of questions between 10 and 200 calculated using a sample with 1 million repetitions and showing the time required for its calculation. Case where there is doubt between three different answers.

Num. Quest.	Sampling	Sampling
	% Negative	Time (s)
10	29.92%	34,366.795
20	15.15%	79,472.383
30	16.68%	125,919.821
40	9.66%	169,065.272
50	10.39%	156,900.51
60	6.25%	251,564.172
70	6.68%	296,882.363
80	4.17%	344,317.804
90	4.43%	315,948.558
100	2.84%	464,915.714
110	2.98%	442,234.988
120	1.88%	492,616.193
130	1.97%	535,646.277
140	1.28%	508,676.051
150	1.36%	521,706.795
160	0.88%	564,736.423
170	0.94%	607,766.221
180	0.62%	650,769.943
190	0.65%	693,826.342
200	0.44%	736,856.236

Figure 2 represents the numerical values contained in Tables 13 and 14. Both tables refer to the case where there is doubt over three different answers. Please also note that the results presented in Table 13 mean that when an examinee has doubts over three different answers in 80 questions and answers all of them at random, the probability of getting a negative score is under 5%. As in the case of Figure 1, the percentage of cases in which negative scores are reached decreases rapidly as the number of questions increases. As also seen in Figure 1, it can be seen that it is not always true that this percentage, although it decreases, is lower for a number of questions $n + 1$ when compared to n, but rather that if n is a multiple of 4, for $n + 1$, the percentage of cases in which a negative score would be reached becomes greater than for n.

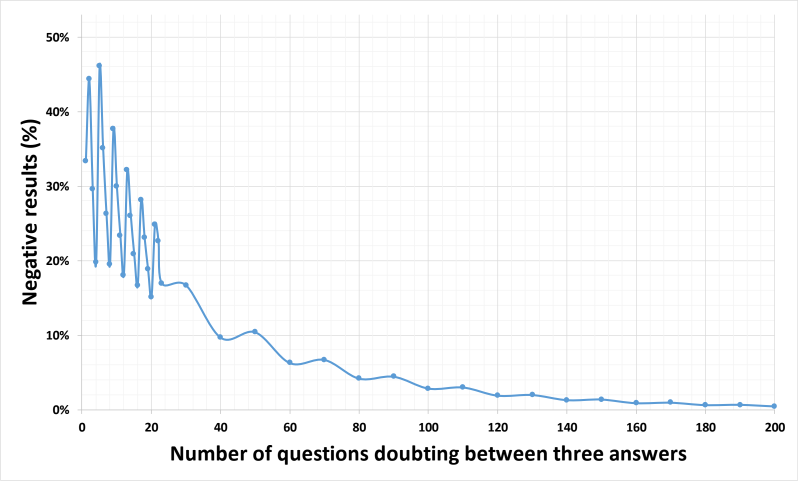

Figure 2. Percentage of negative results when doubting between three answers in a number of questions from 1 to 100.

Table 14. Simulation of the results that would be obtained when carrying out a million tests of 200 questions in which all three answer options are doubted over.

Points	N	Accumulated	Accumulated Percentage
10	15,209	1.5209%	10
20	32,286	3.2286%	20
30	83,582	8.3582%	30
40	140,326	14.0326%	40
50	267,779	26.7779%	50
60	375,425	37.5425%	60
70	554,013	55.4013%	70
80	667,284	66.7284%	80
90	810,408	81.0408%	90
100	879,820	87.9820%	100
110	946,652	94.6652%	110
120	971,724	97.1724%	120
130	990,414	99.0414%	130
140	995,670	99.5670%	140
150	998,870	99.8870%	150
160	999,587	99.9587%	160
170	999,931	99.9931%	170
180	999,973	99.9973%	180
190	999,993	99.9993%	190
200	999,998	99.9998%	200
210	1,000,000	100.0000%	210
220	1,000,000	100.0000%	220

In the case of the MIR test, it is of interest to consider, for example, what the probability is of exceeding the cut-off mark when doubting between three answers in all the exam questions. In order to answer these questions, a simulation of the scores that would be obtained for a test with 200 questions in which each correct question is valued with three points and one point is subtracted for each failed question has been performed. Such a simulation was performed 1 million times. The source code employed is detailed in Table A6 of Appendix A. Thus, for example, given the information in the aforementioned table, if one takes into account that in 55.4013% of the repetitions of the test the score obtained is equal to or less than 70 points, in 44.5987% of the remaining cases, the score obtained will be over 70 points. However, only in 0.0002% of cases would the score be greater than 200, and, therefore, the cut-off mark would be exceeded in only 200 cases out of every million. In a new scenario with the future change of the cut-off mark, the probability of obtaining a score higher than 150 when doubting over all the questions among three options is 0.113%.

As can be observed in all the results presented in this section, the same values have been achieved regardless of the method employed. Therefore, the validation of the results obtained is two-fold. On the one hand, all the methodologies presented are described in detail, and, on the other hand, the coincidences of the results obtained by applying different methodologies make us suppose that they are correct.

4. Conclusions

The present research has achieved its intended aim, as it has been possible to analyze how answering multiple-choice test questions at random affects the final score of examinees, making use of different approaches. It must also be remarked that the authors have developed three new methodologies that may be considered as three different approaches to the same problem.

The main conclusion of the present research is that given that the penalty applied to the questions of the MIR test is calculated so that someone who randomly answers all the questions obtains a zero in the test, it is advisable to answer all of them even if it is not known which the correct answer is. In addition, the results obtained in this work prove that in the case of doubting over two or three of the four possible answers in a certain group of n questions, answering them will, in most cases, yield a net positive result. Moreover, it should be noted that in the case of doubting all the questions of the MIR test between two answer options to each of the questions, in about 50% of the cases, it would be possible to exceed the cut-off mark. With the new cut-off score calculation, this probability is greater than 95%. Finally, regarding the limitations of the present study, it must be considered that one of the main limitations is that the theory developed in this research is only partially presented, focusing more on the computational simulation of the results than on providing a strong theoretical background with theorems and demonstrations. This approach was the most appropriate for this work, since a rapid study of the problem was needed for its application in practice, leaving a deeper mathematical development for future works.

Author Contributions: Conceptualization, J.C. and J.B.R.; methodology, F.S.L., J.C. and P.J.F.; software, F.S.L. and C.P.S.C.; validation, A.G.G., C.P.S.C. and T.V.; formal analysis, T.V. and P.J.F.; investigation, J.C.; resources, J.B.R.; writing—original draft preparation, F.S.L., A.G.G. and T.V.; writing—review and editing, T.V. and C.P.S.C.; visualization, P.J.F.; project administration, J.B.R. and A.G.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Anthony Ashworth for his revision of the English grammar and spelling in the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Source Code

Table A1. Source code that calculates for any number of questions between 1 and 200 the number of cases in which the total score obtained is negative, null and positive when doubting between two answer options.

```
library(tictoc)
library(gtools)
sum_vector<-0
x <- c(3,-1)
for (k in 1:200) {
    tic()
    permut_vector<-permutations(n= 2, r= k, v= x, repeats.allowed = TRUE)
    sum_vector<-0
    for (i in 1:nrow(permut_vector)) {
        sum_vector<-cbind(sum_vector,sum(permut_vector[i,]))
    }
    sum_vector<-sum_vector[-1]
    print(
        paste( k, ": ", "negative:", sum(sum_vector<0), "zero:", sum(sum_vector==0), "positive:",
sum(sum_vector>0), "total:",nrow(permut_vector) )
    )
    toc()
}
```

Table A2. Source code that calculates for any number of questions between 2 and 200, by means of a sample of 1 million replicas, the number of cases in which the total score obtained is negative, null and positive.

```
library(tictoc)
x <- c(3,-1)
for (k in 2:200) {
    tic()
    sum_vector<-0
    list_of_results<-0
    for (index in 1:1E6) {
        results<-sample(x,k,replace=TRUE)
        list_of_results<-rbind(list_of_results,results)
    }
    list_of_results<-list_of_results[-1,]
    sum_vector<-rowSums(list_of_results)
    sum_vector<-sum_vector[-1]
    print(paste(k,": ",sum(sum_vector<0)/nrow(list_of_results)*100))
    toc()
}
```

Table A3. Source code that simulates the results that would be obtained when carrying out a million tests of 200 questions in which all of those questions are doubted over between two answer options.

```
library(tictoc)
x <- c(3,-1)
k<-200
tic()
sum_vector<-0
list_of_results<-0
for (indice in 1:1E6) {
   resultados<-sample(x,k,replace=TRUE)
   list_of_results<-rbind(list_of_results,results)
}
list_of_results<-list_of_results[-1,]
sum_vector<-rowSums(list_of_results)
print(paste(k,": ",sum(sum_vector<0)/nrow(list_of_results)*100,sum(sum_vector>k*3/3)/
nrow(list_of_results)*100))
toc()
```

Table A4. Source code that calculates for any number of questions between 1 and 200 the number of cases in which the total score obtained is negative, null and positive when doubting over three answer options.

```
library(tictoc)
library(gtools)
sum_vector<-0
x <- c('i1','i2','c')
k<-3
for (k in 1:200) {
   tic()
   permut_vector<-permutations(n= 3, r= k, v= x, repeats.allowed = TRUE)
   permut_vector<-matrix(as.numeric(gsub("c", 3, gsub("i1", -1,gsub("i2", -1,
permut_vector)))),nrow=3ˆk)
   sum_vector<-0
   for (i in 1:nrow(permut_vector)) {
      sumas<-cbind(sum_vector,sum(permut_vector[i,]))
   }
   sum_vector<-sum_vector[-1]
   print(paste(k,": ",sum(sum_vector<0),sum(sum_vector==0),sum(sum_vector>0),
nrow(permut_vector)))
   toc()
}
```

Table A5. Source code that calculates for any number of questions between 2 and 200, by means of a sample of 1 million replicas, the number of cases in which the total score obtained is negative, null and positive, taking into account that there are doubts over 3 possible answers to each question.

```
library(tictoc)
x <- c(3,-1)
for (k in 2:200) {
  tic()
  sum_vector<-0
  lista_of_results<-0
  for (index in 1:1E6) {
    results<-sample(x,k,prob=c(1/3,2/3),replace=TRUE)
    lista_of_results<-rbind(lista_of_results,results)
  }
  list_of_results<-list_of_results[-1,]
  sum_vector<-rowSums(list_of_results)
  sum_vector<-sum_vector[-1]
  print(paste(k,": ",sum(sum_vector<0)/nrow(list_of_results)*100))
  toc()
}
```

Table A6. Source code that simulates the results that would be obtained when carrying out a million tests of 200 questions in which all of those questions are doubted over among three answer options.

```
library(tictoc)
x <- c(3,-1,-1)
k<-200
tic()
sum_vector<-0
list_of_results<-0
for (index in 1:1E6) {
  results<-sample(x,k,replace=TRUE)
  list_of_results<-rbind(list_of_results,results)
}
list_of_results<-list_of_results[-1,]
sumas<-rowSums(list_of_results)
print(paste(k,": ",sum(sum_vector<0)/nrow(list_of_results)*100,sum(sum_vector>k*3/3)/
nrow(list_of_results)*100))
toc()
```

References

1. Paludan, A. *Chronicle of the Chinese Emperors: The Reign-by-Reign Record of the Rulers of Imperial China*; Thames and Hudson: New York, NY, USA, 1998.
2. Binet, A.; Simon, T. The development of the Binet-Simon scale, 1905–1908. In *Readings in the History of Psychology*; Dennis, W., Ed.; Appleton-Century-Crofts, Inc.: New York, NY, USA, 1948; pp. 412–424. [CrossRef]
3. Traub, R.E. Classical Test Theory in Historical Perspective. *Educ. Meas. Issues Pract.* **2005**, *16*, 8–14. [CrossRef]
4. Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1980.
5. Marsh, E.J.; Roediger, H.L.; Bjork, R.A.; Bjork, E.L. The memorial consequences of multiple-choice testing. *Psychon. Bull. Rev.* **2007**, *14*, 194–199. [CrossRef] [PubMed]
6. Hasher, L.; Goldstein, D.; Toppino, T. Frequency and the conference of referential validity. *J. Verbal Learn. Verbal Behav.* **1977**, *16*, 107–112. [CrossRef]
7. Baladrón, J.; Sánchez Lasheras, F.; Romeo Ladrero, J.M.; Villacampa, T.; Curbelo, J.; Jiménez Fonseca, P.; García Guerrero, A. The MIR 2018 Exam: Psychometric Study and Comparison with the Previous Nine Years. *Medicina* **2019**, *55*, 751. [CrossRef] [PubMed]
8. Budescu, D.; Bar-Hillel, M. To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring. *J. Educ. Meas.* **1993**, *30*, 277–291. [CrossRef]
9. Espinosa, M.P.; Gardeazabal, J. Optimal correction for guessing in multiple-choice tests. *J. Math. Psychol.* **2010**, *54*, 415–425. [CrossRef]
10. Bliss, L.B. A Test of Lord's Assumption regarding Examinee Guessing Behavior on Multiple-Choice Tests Using Elementary School Students. *J. Educ. Meas.* **1980**, *17*, 147–153. Available online: http://www.jstor.org/stable/1434807 (accessed on 15 September 2022). [CrossRef]

11. Higham, P.A.; Arnold, M.M. How many questions should I answer? Using bias profiles to estimate optimal bias and maximum score on formula-scored tests. *Eur. J. Cogn. Psychol.* **2007**, *19*, 718–742. [CrossRef]
12. Catanzano, T.; Jordan, S.G.; Lewis, P.J. Great Question! The Art and Science of Crafting High-Quality Multiple-Choice Questions. *J. Am. Coll. Radiol.* **2022**, *19*, 687–692. [CrossRef] [PubMed]
13. Rodriguez-Torrealba, R.; Garcia-Lopez, E.; Garcia-Cabot, A. End-to-End generation of Multiple-Choice questions using Text-to-Text transfer Transformer models. *Expert Syst. Appl.* **2022**, *208*, 118258. [CrossRef]
14. Haladyna, T.M.; Downing, S.M. A Taxonomy of Multiple-Choice Item-Writing Rules. *Appl. Meas. Educ.* **1989**, *2*, 37–50. [CrossRef]
15. Danh, T.; Desiderio, T.; Herrmann, V.; Lyons, H.M.; Patrick, F.; Wantuch, G.A.; Dell, K.A. Evaluating the quality of multiple-choice questions in a NAPLEX preparation book. *Curr. Pharm. Teach. Learn.* **2020**, *12*, 1188–1193. [CrossRef] [PubMed]
16. Coughlin, P.A.; Featherstone, C.R. How to Write a High Quality Multiple Choice Question (MCQ): A Guide for Clinicians. *Eur. J. Vasc. Endovasc. Surg.* **2017**, *54*, 654–658. [CrossRef] [PubMed]
17. Al Muhaissen, S.A.; Ratka, A.; Akour, A.; Alkhatib, H.S. Quantitative analysis of single best answer multiple choice questions in pharmaceutics. *Curr. Pharm. Teach. Learn.* **2019**, *11*, 251–257. [CrossRef] [PubMed]
18. Kumar, D.; Jaipurkar, R.; Shekhar, A.; Sikri, G.; Srinivas, V. Item analysis of multiple choice questions: A quality assurance test for an assessment tool. *Med. J. Armed Forces India* **2021**, *77* (Suppl. S1), 85–89. [CrossRef] [PubMed]
19. Burton, R.F. Quantifying the Effects of Chance in Multiple Choice and True/False Tests: Question selection and guessing of answers. *Assess. Eval. High. Educ.* **2001**, *26*, 41–50. [CrossRef]
20. Şad, S.N. Does difficulty-based item order matter in multiple-choice exams? (Empirical evidence from university students). *Stud. Educ. Eval.* **2020**, *64*, 100812. [CrossRef]
21. Wang, C.-N.; Yang, F.-C.; Nguyen, V.T.T.; Nguyen, Q.M.; Huynh, N.T.; Huynh, T.T. Optimal Design for Compliant Mechanism Flexure Hinges: Bridge-Type. *Micromachines* **2021**, *12*, 1304. [CrossRef] [PubMed]
22. Wang, C.-N.; Yang, F.-C.; Nguyen, V.T.T.; Vo, N.T.M. CFD Analysis and Optimum Design for a Centrifugal Pump Using an Effectively Artificial Intelligent Algorithm. *Micromachines* **2022**, *13*, 1208. [CrossRef] [PubMed]
23. Nguyen, T.V.T.; Huynh, N.-T.; Vu, N.-C.; Kieu, V.N.D.; Huang, S.-C. Optimizing compliant gripper mechanism design by employing an effective bi-algorithm: Fuzzy logic and ANFIS. *Microsyst. Technol.* **2021**, *27*, 3389–3412. [CrossRef]
24. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022. Available online: https://www.R-project.org/ (accessed on 15 September 2022).