Analysis of the reliability of multiple-choice tests with the Monte Carlo method

Estudio de la fiabilidad de test multirrespuesta con el método de Monte Carlo

DOI: 10.4438/1988-592X-RE-2021-392-479

José Calaf Chica
María José García Tárrago
Universidad de Burgos

Abstract

During the twentieth century, there has been a lot published about the reliability of the multiple-choice tests for subject evaluation. Specifically, there are many theoretical and empirical studies that compare the different scoring methods applied in tests. In this study, a novel algorithm was designed to generate hypothetical examinees with three specific characteristics: real knowledge, level of cautiousness and erroneous knowledge. The first characteristic established the probability of a student to knowing the veracity or falsity of each choice in a multiple-choice test. The level of cautiousness showed the probability of answering a question not known by guessing. Finally, erroneous knowledge was false knowledge assimilated as being true. The test setup required by the algorithm included the test length, choices per question and the scoring system. The algorithm sent tests to these hypothetical examinees analyzing the deviation between the real knowledge and the estimated knowledge (the test score reached). The most popular test scoring methods (positive marking, negative marking, free-choice tests and the dual response method) were analyzed and compared to measure their reliability. To validate the algorithm, this was compared with an analytical probabilistic model. This study verified that the presence of erroneous knowledge or lack there of generates an important alteration in the reliability of the most accepted scoring methods in the educational community (the negative marking method). Given the impossibility of ascertaining the existence of

Received: 29-05-2020 Accepted: 27-11-2020

erroneous knowledge in the examinees using a test, it is up to the examiner whether or not to penalize the presence of such knowledge with the use of negative marking or to find a closer estimation of the real knowledge using the positive marking method.

Keywords: Multiple Choice Test, Computer Simulation, Scoring, Evaluation, Monte Carlo Method.

Resumen

Durante gran parte del siglo XX se ha escrito mucho sobre la fiabilidad de los test multirrespuesta como método para la evaluación de contenidos. En concreto son muchos los estudios teóricos y empíricos que buscan enfrentar los distintos sistemas de puntuación existentes. En esta investigación se ha diseñado un algoritmo que genera estudiantes virtuales con los siguientes atributos: conocimiento real, nivel de cautela y conocimiento erróneo. El primer parámetro establece la probabilidad que tiene el alumno de conocer la veracidad o falsedad de cada opción de respuesta del test. El nivel de cautela refleja la probabilidad de responder a una cuestión desconocida. Finalmente, el conocimiento erróneo es aquel conocimiento falsamente asimilado como cierto. El algoritmo también tiene en cuenta parámetros de configuración del test como el número de preguntas, el número de opciones de respuesta por pregunta y el sistema de puntuación establecido. El algoritmo lanza test a los individuos virtuales analizando la desviación generada entre el conocimiento real y el conocimiento estimado (la puntuación alcanzada en el test). En este estudio se confrontaron los sistemas de puntuación más comúnmente utilizados (marcado positivo, marcado negativo, test de elección libre y método de la respuesta doble) para comprobar la fiabilidad de cada uno de ellos. Para la validación del algoritmo, se comparó con un modelo analítico probabilístico. De los resultados obtenidos, se observó que la existencia o no de conocimiento erróneo generaba una importante alteración en la fiabilidad de los test más aceptados por la comunidad educativa (los test de marcado negativo). Ante la imposibilidad de comprobar la existencia de conocimiento erróneo en los individuos a través de un test, es decisión del evaluador castigar su presencia con el uso del marcado negativo, o buscar una estimación más real del conocimiento real a través del marcado positivo.

Palabras clave: Test Multirrespuesta, Simulación Computacional, Puntuación, Evaluación, Método de Monte Carlo.

Received: 29-05-2020 Accepted: 27-11-2020

Introduction

Multiple-choice tests have been widely applied in the majority of stages of the educational system in many countries. Even the certification of competencies or skills in many areas of the industrial or medical sectors are often based on this method of evaluation. They provide an interesting tool when a high number of examinees must be evaluated. The reliability of the method for grading is crucial when there is a passing mark which defines the pass/fail threshold in the examinee certification or graduation. This has been the prime motivation for research publications and investigation related to multiple-choice tests. (Papenberg, Diedenhofen, and Musch 2019; Parkes and Zimmaro 2016). In these methods of evaluation, an assertion, also called the stem of the question, is introduced and the examinee has to choose one of a selection of multiple answers, where one of them is the key, the correct option, and the others are distractors, the wrong answers. An important point in any research about the reliability of the multiple-choice tests is that distractors have to be well-designed (Burton 2005; Hsu et al. 2018). It means that the falsity of the distractor should be only clear to an examinee who knows the evaluated subject in that question.

Beyond this typological classification, there is a wide list of alternatives to mark or evaluate the tests after they are filled in by the examinees. The simplest way is the 'Number Right' method (Kurz 1999) where the selection of a correct answer (the key answer) is registered and marked with a positive value, and the selection of any distractor or unanswered questions implies no score for that question. The main problem of this scoring method is the deviation generated between the real knowledge and the estimated knowledge of the examinees due to guessing. The student, after having marked the questions that he/she knows, tends to guess the remaining questions, since the selection of distractors does not imply any penalization or negative score (Lin 2018). A way to reduce the bias in the real knowledge versus the estimated knowledge generated by this evaluation method was achieved with the 'Negative Marking' method. In this scoring system, the selection of any distractor is scored with a negative mark, so any mistakes are penalized, and the examinees are dissuaded from guessing. However, there is another motivation for selecting the distractors: erroneous knowledge (Burton 2004). This is false knowledge assimilated as true by the examinee. Thus, the incorrectly selected answers would come from a combination of this erroneous knowledge and guessing, concluding that Negative Marking penalizes both behaviors equally. It is not possible to quantify or discern the relative weight of each one in the final score of the test. Considering that the essence of the Negative Marking is based on the elimination of guessing by means of a penalty gauged to reduce its influence in the final score, the presence of erroneous knowledge would reduce and hide the real knowledge of the examinee. Therefore, estimated knowledge could be lower than real knowledge.

The specific value of the scoring penalty that should be imposed for each distractor selection in the Negative Marking method is generally established using the probability theory to reach the null expected value by guessing (Warwick, Bush, and Jennings 2010). The calculation of this sanction values is based on the equation (1):

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

where p is the value of the penalty and k the number of answers per question.

A new concept, 'partial knowledge' (Slepkov and Godfrey 2019), leads to an interesting matter that should be included in this discussion. This is based more on an examinee's behavior than in knowledge typology. It is defined as the capability of the examinees to discern some but not all of the distractors in a question (Betts et al. 2009). This reduces the remaining choices of the question and, in a scenario of guessing, the probability of answering correctly increases significantly, with no alterations in the sanction value by guessing. Considering that sanction established in a Negative Marking method is fixed with the equation (1), which considers a k number of answers, the expected value in the final score (the estimated knowledge) would be higher than the real knowledge of the examinee (Budescu and Bar-Hillel 1993).

Other parameters influence the reliability of this evaluation method. Specifically, the test length or the number of questions considered in the test. To ensure the validity of equation (1), the test length must be great enough to ensure a minimal scattering in the correlation between the estimated and real knowledge. The value of the penalty is based on probability theory and, in consequence, needs enough random items to work properly. But a sufficient number of random items is not proportional to the number of questions, because an examinee with greater knowledge answers few questions by guessing compared to a lower-knowledge examinee. The expected value established by the penalty is easily obtained for lower-knowledge scenarios and extremely difficult in higher-knowledge scenarios. This implies that the knowledge level of an examinee influences the validity of equation (1) or the reliability of tests that use the Negative Marking scoring method.

For the moment, it has been suggested that an individual, at the moment when he/she does not know the correct answer for a question, tries to respond by guessing. But analyzing the individual's behavior, this assertion would be inappropriate. In reality, it is more complex, and this is where a new parameter comes in: the level of cautiousness of each examinee (Espinosa and Gardeazabal 2010; Riener and Wagner 2017). The sanction in the Negative Marking was used to remove or avoid guessing, but not all examinees can be said to have the same degree of bold or cautious behavior (Moon, Keehner, and Katz 2020). In addition, this influence depends on the personality of each individual, and over-cautiousness is a variable independent of real knowledge (Hammond et al. 1998). In Negative Marking, an over-cautiousness examinee is more greatly influenced by the threat of a sanction than bold examinees. The number of unmarked questions is higher in cautious individuals, and the number of questions answered by guessing is higher in more daring individuals. Therefore, two examinees with different levels of cautiousness but with similar levels of real knowledge would show different test scores and estimated knowledge. In addition, the bolder examinees make the most of partial knowledge to improve their final score, because the probability of selecting the right answer is higher than the probability used to calculate the penalty. The over-cautious examinees only answer the question when they know the correct option, so they never make the most of those opportunities to improve their final score.

In this investigation, four scoring methods have been used: the preceding two methods, the Number Right and the Negative Marking, the "Free Choice" method (Jennings and Bush 2006) and the "Elimination Procedure" method (Bush 2015). The Free Choice method can be distinguished by allowing a selection of multiple answers. The motivation for the implementation of this rule is based on the rewarding of the examinees' partial knowledge. For example, in a four-choices test, this scoring system would work as follows: If the right answer is marked, the examinee is rewarded with one point (3/3); if the examinee marks the right one and one distractor, he/she is rewarded with (3-1)/3 = 0.67 points; finally, if three items are selected (the right one and two distractors) the reward is equal to (3-2)/3 = 0.33 points. When the examinee does not select the right answer, he/she is punished with: -0.33 points (one distractor selected), -0.67 points (two distractors selected), and -1 point (three distractors selected). There is an alternative method, similar to the 'Free-Choice' test, called the 'Dual Response' system (Akeroyd 1982). In this scoring method, multiple-answers selection is also allowed, but the rewarding system changes: one point if only the right answer is selected, 0.5 points for selecting the right one and one distractor, 0.25 points for the right one and two distractors and no points or penalties for other selections. The Elimination Procedure scoring method is similar to the Free-Choice method in the sense that multiple answers can be selected in the same question, but, in this "elimination procedure", the examinee has to select the distractors instead of the right answer.

This introduction shows the underlying complexity of the analysis of each scoring method, mostly in the empirical studies where variables like cautiousness, erroneous knowledge or partial knowledge influence the final score, and it is not possible to discern their existence or to quantify their relative weight. Analytical studies using the probability theory as an alternative could become complex if the analysis considers all the variables introduced previously. That is the reason why this investigation reflected on the possibility of using the potentiality of computer algorithms. The main objective was the design of a code to generate hypothetical examinees characterized by different input parameters (real knowledge, erroneous knowledge and level of cautiousness). Combining this database of hypothetical examinees with different test designs, the algorithm would give, as an output, the final score or estimated knowledge for each examinee. This system would make it possible to interpret parameters that are difficult or impossible to analyze using empirical research, since cautiousness influences the final score.

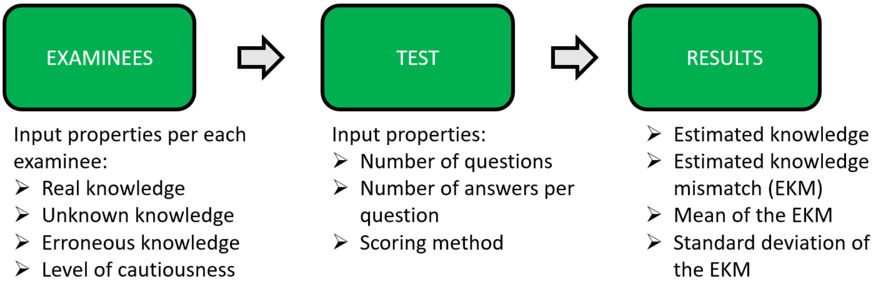
Received: 29-05-2020 Accepted: 27-11-2020

Methodology

The main objective of this investigation was the development of an algorithm to simulate the filling in of a multiple-choice test. Python was the programming language selected to develop the code because of its simplicity, capability, readability and extensive libraries with built-in modules.

Figure I shows a basic flowchart of the algorithm. There are three main blocks: the examinees, the test and the results. Examinees and tests have different input parameters that define their properties. The results block is related to the output data obtained by the algorithm. Each one of these input and output parameters is defined in the following sections.

FIGURE I. Basic flowchart of the algorithm



Input properties per each examinee:
- Real knowledge
- Unknown knowledge
- Erroneous knowledge
- Level of cautiousness

Input properties:
- Number of questions
- Number of answers per question
- Scoring method

- Estimated knowledge
- Estimated knowledge mismatch (EKM)
- Mean of the EKM
- Standard deviation of the EKM

Examinees properties

This algorithm measures a hypothetical 'Subject Knowledge' introduced by the test, where each examinee has assimilated a specific percentage of that Subject Knowledge (called 'Real Knowledge' of the examinee). Figure II shows a schematic view of the Subject Knowledge (blue rectangle) which is classified as 'Real Knowledge' (the knowledge assimilated by each examinee) seen as the green rectangle and the rest of the Subject Knowledge, called 'Lack of Knowledge'. Figure II also shows the classification established for this Lack of Knowledge, in which there is 'Unknown Knowledge' and 'Erroneous Knowledge'. Unknown

Knowledge is the Subject Knowledge that the examinee has not retained, and Erroneous Knowledge is the percentage of the Lack of Knowledge that has been misunderstood. Figure III shows an examinee's knowledge classification using an example of managing addition equations. The Real Knowledge would be related to the correct equations (the examinee knows how to manage some specific addition equations). The Unknown Knowledge would be related to the equations that the examinee does not know how to calculate. Finally, the Erroneous Knowledge would be related to the equations that the examinee believes he/she knows but are actually wrong (misunderstood knowledge).

FIGURE II. Classification of the Subject Knowledge

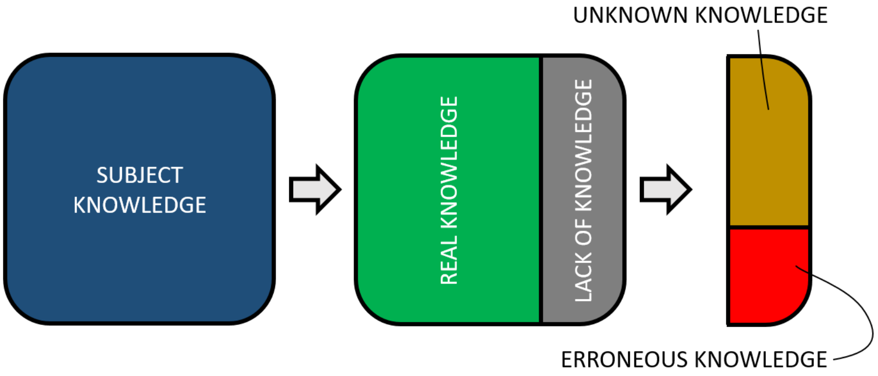
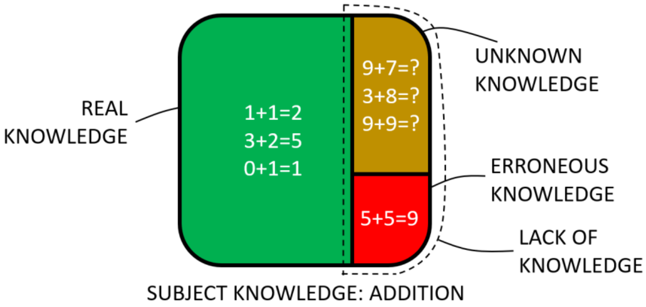


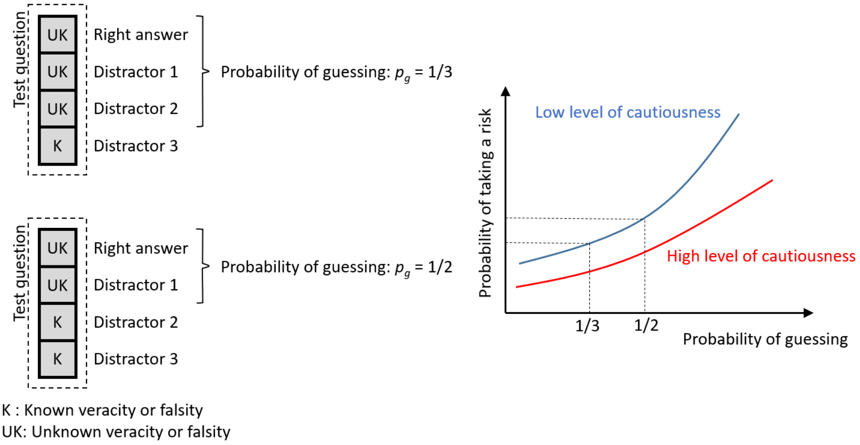
FIGURE III. Example of the interpretation of the knowledge classification



Received: 29-05-2020 Accepted: 27-11-2020

Another property for each examinee introduced in the input parameters of the algorithm was the Level of Cautiousness. It measures the examinee's capability to take a risk and try to guess a test question that he/she does not know the right answer to. However, it is important to clarify that the probability of taking a risk does not only depend on the examinee's level of cautiousness. The probability of guessing also influences the probability of taking a risk. Figure IV shows a schematic example of two four-choice questions. In these examples, the first answer option is the right answer or key answer, and the rest of the answer options are the distractors. For the first case in Figure IV (UK,UK,UK,K), the examinee would not know the right answer (identified with UK), would not know two distractors (also identified with UK) and only would know (identified with K) the falsity of the last distractor. Thus, the probability of guessing would be equal to p_g*=1/3*. If the examinee knew two distractors, not knowing one distractor and the right answer, the probability of guessing would grow to p_g*=1/2*. The graph included in Figure IV shows that an increment in the probability of guessing increases the probability of taking a risk. Thus, there is a linear relationship between the probability of guessing and the probability of taking a risk. In addition, there are as many linear relations as levels of cautiousness defined. The methodology followed in the algorithm to simulate these curves is included in the Annex.

FIGURE IV. Dependency of the probability of taking a risk on the level of cautiousness and the probability of guessing.


Test question
UK
UK
UK
K
Right answer
Distractor 1
Distractor 2
Distractor 3
Probability of guessing: p_g = 1/3
Test question
UK
UK
K
K
Right answer
Distractor 1
Distractor 2
Distractor 3
Probability of guessing: p_g = 1/2
Low level of cautiousness
High level of cautiousness
Probability of taking a risk
1/3
1/2
Probability of guessing
K : Known veracity or falsity
UK: Unknown veracity or falsity

Test properties

Three test properties must be introduced in the algorithm as input parameters: the length of the test, the number of answers per question and the scoring method. The length of the test measures the number of questions evaluated in the test. The number of answers per question is related to the number of distractors associated with the right answer. Finally, the scoring method (Number Right, Negative Marking, Free-Choice and Dual Response) is the method used to rate each question. All of them are explained in the introduction of this investigation. For the Negative Marking and Free-Choice methods, the penalty value would also be an input parameter of the test.

Results output

The algorithm calculates the final score of the test for each examinee. Each final score, called 'Estimated Knowledge', is compared with the 'Real Knowledge' of the examinee, obtaining the 'Estimated Knowledge Mismatch' (*EKM*), which is equal to the difference between the Estimated Knowledge and the Real Knowledge (see equation (2)).

$$EKM = EK - RK \tag{2}$$

where *RK* is the Real Knowledge and *EK* is the Estimated Knowledge.

The algorithm obtains an *EKM* for each examinee, and it also calculates the mean value μ_{EKM} and the standard deviation σ_{EKM} of the *EKM*s (see equation (3)).

$$\mu_{EKM} = \frac{\sum_{i=1}^{n} EKM_i}{n} \qquad \sigma_{EKM} = \sqrt{\frac{\sum_{i=1}^{n}[EKM_i - \mu_{EKM}]^2}{n}} \tag{3}$$

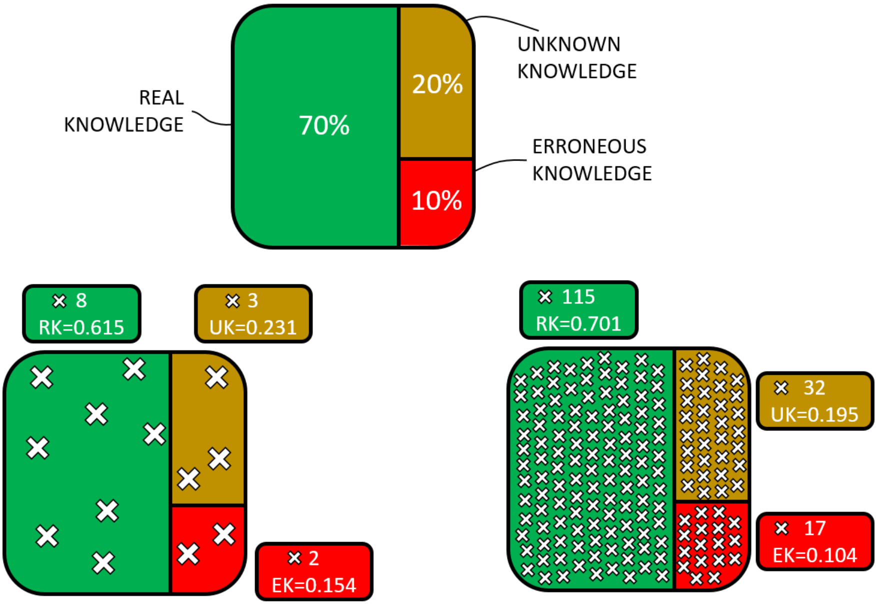
where *EKM* is the Estimated Knowledge Mismatch and n is the number of evaluated examinees.

Received: 29-05-2020 Accepted: 27-11-2020

Simulation of the examinee/questions interaction

The algorithm uses the Monte Carlo method to generate the interaction between the examinees and the test questions. Figure V shows how the algorithm applies this statistical method. A random function is launched for each answer option of a test question. The random value obtained by this random function (represented as a white cross in Figure V) can fall in the region of the Real Knowledge, the Unknown Knowledge or the Erroneous Knowledge of the examinee. If the Real Knowledge is high enough, the random value will easily fall on it. Thus, if the examinee has a high Real Knowledge, he/she will generally know the veracity or falsity of the answer options. In the example represented in Figure V, an examinee with a Real Knowledge of 70%, Unknown Knowledge of 20% and Erroneous Knowledge of 10% is represented with blocks. Each block has a corresponding area to its assigned percentage. As mentioned previously, each cross would be a random value and, for the specific case of the algorithm analyzed in this investigation, every cross would be one attempt to discern if the examinee does or does not know an answer option for a question. If the attempt falls in the Real Knowledge block, the examinee will know (answer ticked as K) the veracity or falsity of this answer option. If the attempt falls in the Unknown Knowledge block, the examinee will not know (answer ticked as UK) the veracity or falsity of this answer option. Finally, if the attempt falls in the Erroneous Knowledge block, the examinee will confuse a distractor as a right answer and vice versa (answer ticked as EK). Figure V shows that a great quantity of questions in a test (equivalent to a high number of answer options) reduces the difference between the percentages of the different types of knowledge of the examinee and the estimated percentages using the Monte Carlo method.

FIGURE V. Monte Carlo method applied in the analysis of the answer options



When the random function is launched for all the answer options of a four-choice question, the algorithm obtains an identifier like (A_1,A_2,A_3,A_4), where each A_i represents one answer option (A_1 corresponds to the right answer while A_2 to A_4 are the distractors). In each position, the result of using the Monte Carlo method is indicated (K: known answer, UK: unknown answer and EK: erroneously known answer). An example would be (K,UK,UK,UK) where the right answer is known, and all the distractors unknown. In this case, the examinee would tick the correct answer and would receive one point for that question. Another interesting example would be (UK,K,K,EK). In this case, the examinee does not know the right answer, knows two distractors and the third distractor is erroneously known. It means that the examinee would believe that the third distractor is the right answer. Thus, the examinee would tick the third distractor, wrongly answering the question. Another possibility could be (UK,UK,UK,K). It means that the examinee only knows one distractor, not knowing the rest of the answer options, so the examinee could answer the question by guessing. For all the questions in this

Received: 29-05-2020 Accepted: 27-11-2020

situation, the Monte Carlo method would be again launched but using the probability of taking a risk (p_r) by the examinee. This probability p_r is dependent on the level of cautiousness of the examinee and the probability of guessing p_g. The first one is a property of the examinee, and the second one is calculated as $p_g = 1/x$, where x represents the number of answer options not known in the question. Figure VI shows how the probability of guessing and the probability of taking a risk (p_r) are calculated for the specific case (UK,UK,UK,K). The equation of the cautiousness curves represented in Figure VI is included in the Annex.

FIGURE VI. Calculation of the probability of taking a risk

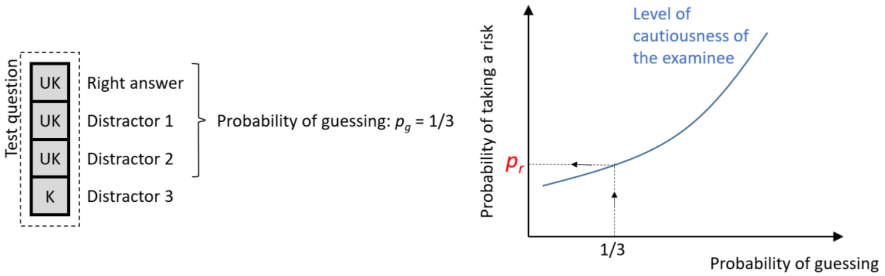
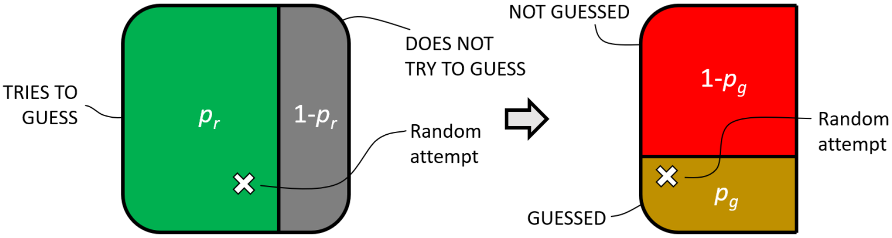

Test question
UK
UK
UK
K
Right answer
Distractor 1
Distractor 2
Distractor 3
Probability of guessing: p_g = 1/3
Probability of taking a risk
Level of cautiousness of the examinee
p_r
1/3
Probability of guessing

Figure VII represents the application of the Monte Carlo method figure out if the examinee does or does not try to guess the right answer to the question and, if the examinee tries, the Monte Carlo method is again applied to perform a random attempt based on the probability of guessing p_g to know if the examinee does or does not guess the right answer.

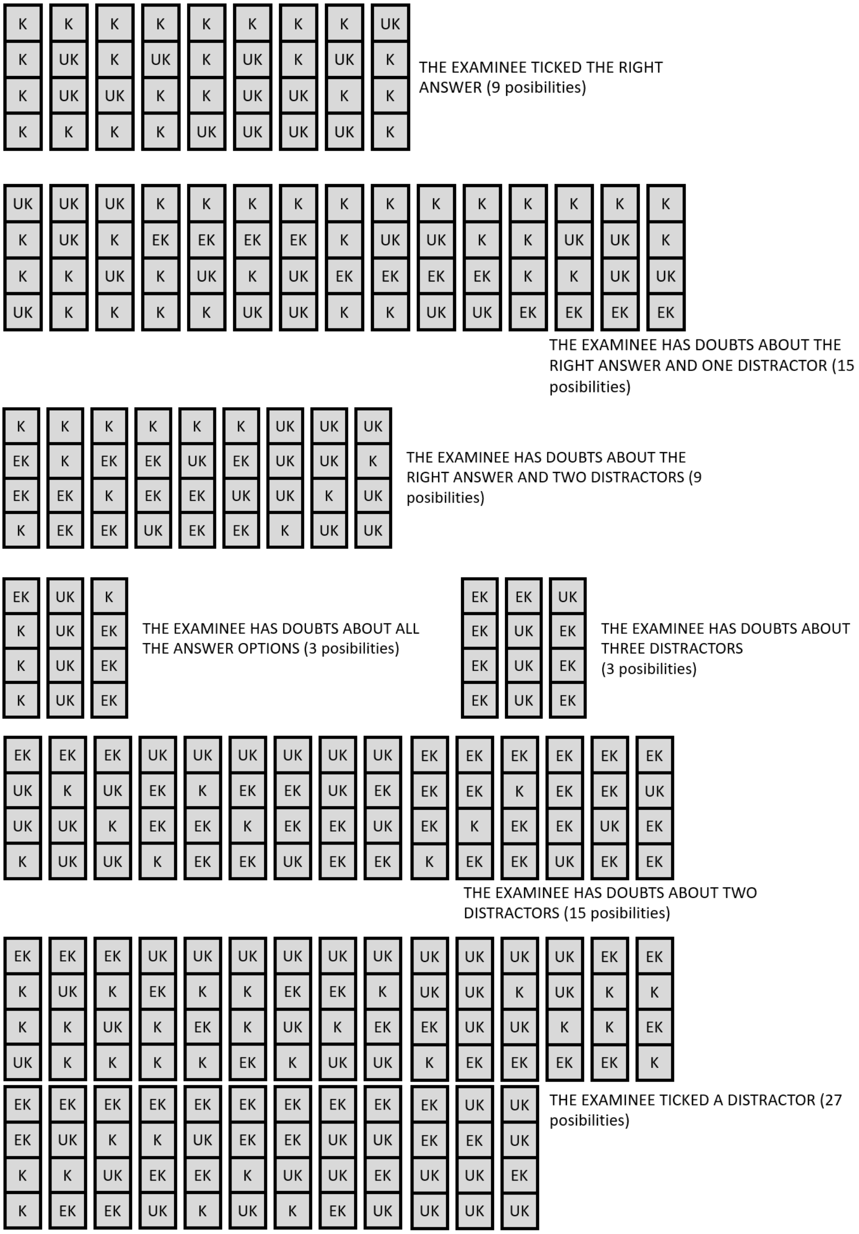
FIGURE VII. Monte Carlo method applied in the analysis of guessing.



Therefore, for a four-choice question, there are $3^4 = 81$ possible combinations of K, UK and EK cases, that is to say a variation of three possibly repeated elements for 4 items per question. Figure VIII shows the examinee behavior for all 81 cases of a four-choice test. For all of the cases in which the examinee had doubts about more than one answer option, the previously explained Monte Carlo method is launched to know if the examinee ticked any answer and, in that case, if the examinee guessed the right answer or not.

Received: 29-05-2020 Accepted: 27-11-2020

FIGURE VIII. Classification of the 81 possibilities in a four-choice question



K	K	K	K	K	K	K	K	UK
K	UK	K	UK	K	UK	K	UK	K
K	UK	UK	K	K	UK	UK	K	K
K	K	K	K	UK	UK	UK	UK	K

THE EXAMINEE TICKED THE RIGHT ANSWER (9 posibilities)

UK	UK	UK	K	K	K	K	K	K	K	K	K	K	K	K
K	UK	K	EK	EK	EK	EK	K	UK	UK	K	K	UK	UK	K
K	K	UK	K	UK	K	UK	EK	EK	EK	EK	K	K	UK	UK
UK	K	K	K	K	UK	UK	K	K	UK	UK	EK	EK	EK	EK

THE EXAMINEE HAS DOUBTS ABOUT THE RIGHT ANSWER AND ONE DISTRACTOR (15 posibilities)

K	K	K	K	K	K	UK	UK	UK
EK	K	EK	EK	UK	EK	UK	UK	K
EK	EK	K	EK	EK	UK	UK	K	UK
K	EK	EK	UK	EK	EK	K	UK	UK

THE EXAMINEE HAS DOUBTS ABOUT THE RIGHT ANSWER AND TWO DISTRACTORS (9 posibilities)

EK	UK	K
K	UK	EK
K	UK	EK
K	UK	EK

THE EXAMINEE HAS DOUBTS ABOUT ALL THE ANSWER OPTIONS (3 posibilities)

EK	EK	UK
EK	UK	EK
EK	UK	EK
EK	UK	EK

THE EXAMINEE HAS DOUBTS ABOUT THREE DISTRACTORS (3 posibilities)

EK	EK	EK	UK	UK	UK	UK	UK	UK	EK	EK	EK	EK	EK	EK
UK	K	UK	EK	K	EK	EK	UK	EK	EK	EK	K	EK	EK	UK
UK	UK	K	EK	EK	K	EK	EK	UK	EK	K	EK	EK	UK	EK
K	UK	UK	K	EK	EK	UK	EK	EK	K	EK	EK	UK	EK	EK

THE EXAMINEE HAS DOUBTS ABOUT TWO DISTRACTORS (15 posibilities)

EK	EK	EK	UK	UK	UK	UK	UK	UK	UK	UK	UK	UK	EK	EK
K	UK	K	EK	K	K	EK	EK	K	UK	UK	K	UK	K	K
K	K	UK	K	EK	K	UK	K	EK	EK	UK	UK	K	K	EK
UK	K	K	K	K	EK	K	UK	UK	K	EK	EK	EK	EK	K

EK	EK	EK	EK	EK	EK	EK	EK	EK	EK	UK	UK
EK	UK	K	K	UK	EK	EK	UK	UK	EK	EK	UK
K	K	UK	EK	EK	K	UK	UK	EK	UK	UK	EK
K	EK	EK	UK	K	UK	K	EK	UK	UK	UK	UK

THE EXAMINEE TICKED A DISTRACTOR (27 posibilities)

Rating assignation process

The algorithm uses four scoring methods: Number Right, Negative Marking, Free-Choice and Dual Response. The process followed by the algorithm to apply each scoring method was previously explained in the Introduction. This process employs the results obtained using the Monte Carlo method. The total or final score obtained in the test is calculated as the sum of all the correctly-answered questions and the difference of all the incorrectly-answered questions (if the scoring method has any criteria for penalties).

Case for validation and systematic analysis

The algorithm was validated by means of a comparison with an analytical model of a simple case using the probability theory. After that, the code was used to analyze the influence of each variable in the mean value of the *EKM*s. This systematic analysis considered 720 cases with the following selection of input variables (all of the questions selected were of the four-choice test type):

- The number of examinees: 1000.
- Length of the test: 10, 20, 30, 40, 70 and 100 questions.
- Real knowledge: low (1), mid-level (2) and high (3).
- Level of cautiousness: low (1), mid-level (2) and high (3).
- Erroneous knowledge: none (1), low (2), mid-level (3) and high (4).
- Scoring method: Number Right (NR), Negative Marking (NM), Free Choice (FC) and Dual Response (DR).

For Real Knowledge, a low level means that the examinees have a Real Knowledge somewhere between 0% and 33% of the Subject Knowledge. The mid-level would be a Real Knowledge from 33% to 66%, and a high level from 66% to 100%. For the Erroneous knowledge, a low level would mean a range of 0 to 33% of the Lack of Knowledge, the mid-level a range of 33% to 66%, and the high level a range of 66% to 100%. A level of 'none' would mean that the examinees have no Erroneous Knowledge. For the Level of Cautiousness, a low level means that the examinees have a level of cautiousness *C* between 0 and 0.33. The mid-level would be a *C* between 0.33 and 0.66, and the high level would be a *C* between 0.66

and 1.0. How this *C* value is used to calculate the probability of taking a risk is explained in the Annex.

Each case was identified with the ID xx-RKx-Cx-xx-EKx. As an example, ID 10-RK1-C1-NR-EK1 represents a test with 10 questions, a low Real Knowledge and low level of cautiousness for the examinees, Number Right scoring method and with no erroneous knowledge.

Results

Case for validation

Before using the algorithm, a simple case that could be analyzed with the probability theory was used to verify the code. Comparing both models, the analytical one and that obtained using the algorithm, the code was verified. The input parameters for this case were:

- The number of examinees: 1000.
- Length of the test: 200 questions.
- Real knowledge: set to 50% for all the examinees.
- Level of cautiousness: not applicable (a Number Right scoring method was used, so the absence of penalties eliminates any sense of danger).
- Erroneous knowledge: none.
- Scoring method: Number Right.

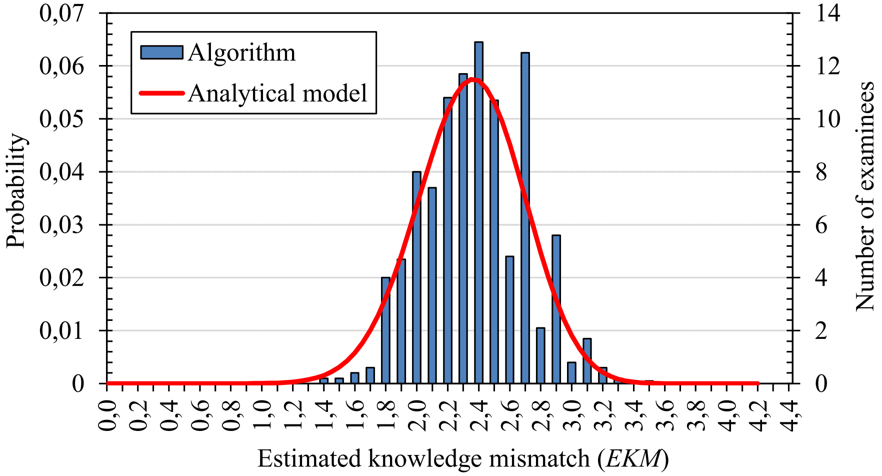
The design of this analytical probability model and the steps followed to calculate the corresponding probability distribution equation are detailed in the Annex. The following equation (4) shows this complex equation. The case for validation analyzed in this investigation used the simplest scoring method (Number Right method) without any erroneous knowledge and no influence of the cautiousness of the examinee. The exponential complexity of the probability model when these input variables are included in the analysis clearly shows the interest and usefulness of an algorithm to simplify and automate this probability calculation.

$$P(S=s)=\sum_{i=0}^{200}\sum_{x=0}^{i}\sum_{j=0}^{200-i}\sum_{y=0}^{j}\sum_{k=0}^{200-j-i}\sum_{z=0}^{k}\{B(200,i,3/16)\cdot B(i,x,1/2)\times \times [B(200,j,3/16)\cdot B(j,y,1/3)]\times[B(200,k,1/16)\cdot B(k,z,1/4)]\} \tag{4}$$

where *x+y+z+[100-(i+j+k)]* must always be equal to *s*.

The following Figure IX shows a graph of the probability of obtaining the different Estimated Knowledge Mismatches (*EKMs*) for equation (4) (red curve). The *EKM* was calculated based on a maximum knowledge of 10. The blue bar chart represents the results registered using the algorithm, in which the examinees are distributed by their different deduced values of *EKM*. A comparison between the probability distribution obtained analytically with the equation (4) and the one obtained using the code developed in this investigation could conclude that the algorithm showed sufficient agreement with the analytical model.

FIGURE IX. Comparison between the analytical model and the result obtained using the algorithm



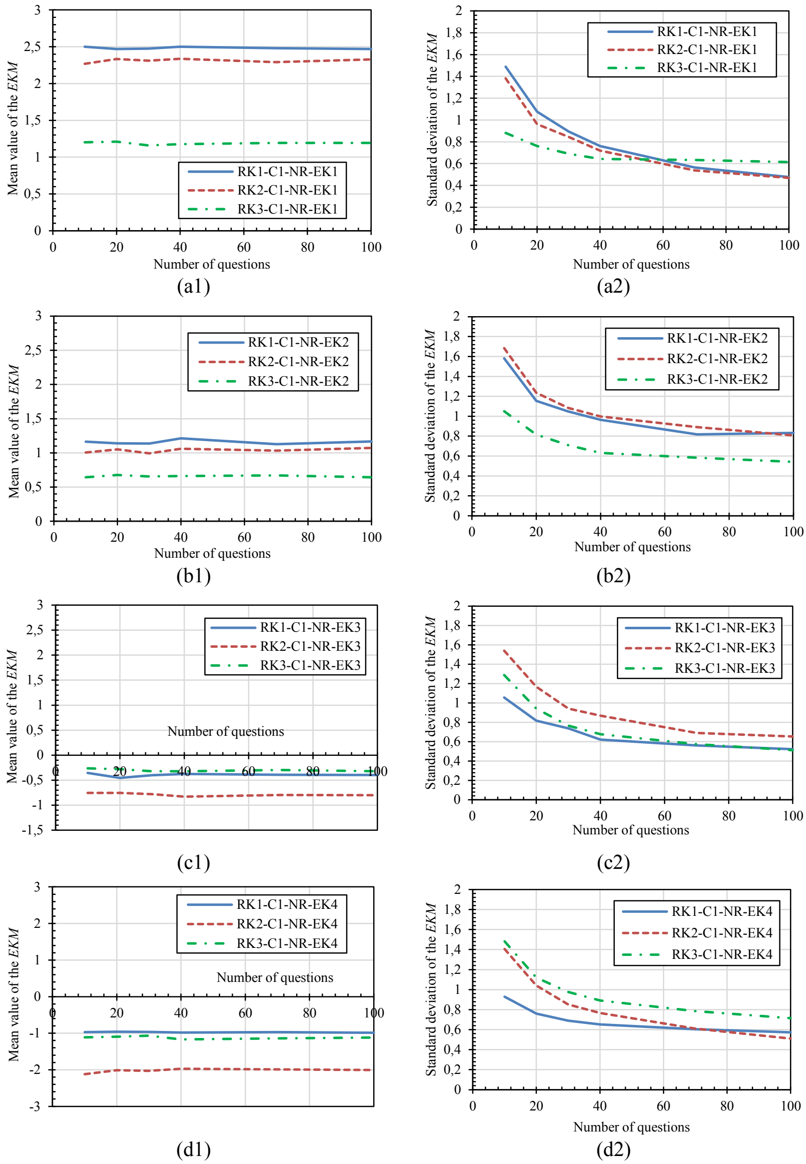
Received: 29-05-2020 Accepted: 27-11-2020

Systematic analysis with the algorithm

As explained in the Methodology section, 720 cases were launched using the algorithm to analyze the influence of each input parameter in the estimated knowledge or final score of a multiple-choice test. One thousand hypothetical examinees were evaluated in each case, obtaining the mean values of the difference between the estimated knowledge and the real knowledge (mean value of the *EKM*s) and the standard deviations of these mean values. The following Figure X shows the mean value of the *EKM*s on the left and the standard deviation on the right, both out of maximum knowledge of 10, versus the number of questions of the tests for the Number Right scoring method. Different levels of Real Knowledge were represented: a blue continuous line for low Real Knowledge (RK1); a red dotted line for mid-level Real Knowledge (RK2); and a dashed line with dots for high Real Knowledge (RK3). From top to bottom, different levels of Erroneous Knowledge are represented: none (EK1), low (EK2), mid-level (EK3), and high (EK4).

Received: 29-05-2020 Accepted: 27-11-2020

FIGURE X. Mean value (1) and standard deviation (2) of the EKM for: (a) none, (b) low, (c) mid-level and (d) high Erroneous Knowledge, using the Number Right method



(a1)

(a2)

(b1)

(b2)

(c1)

(c2)

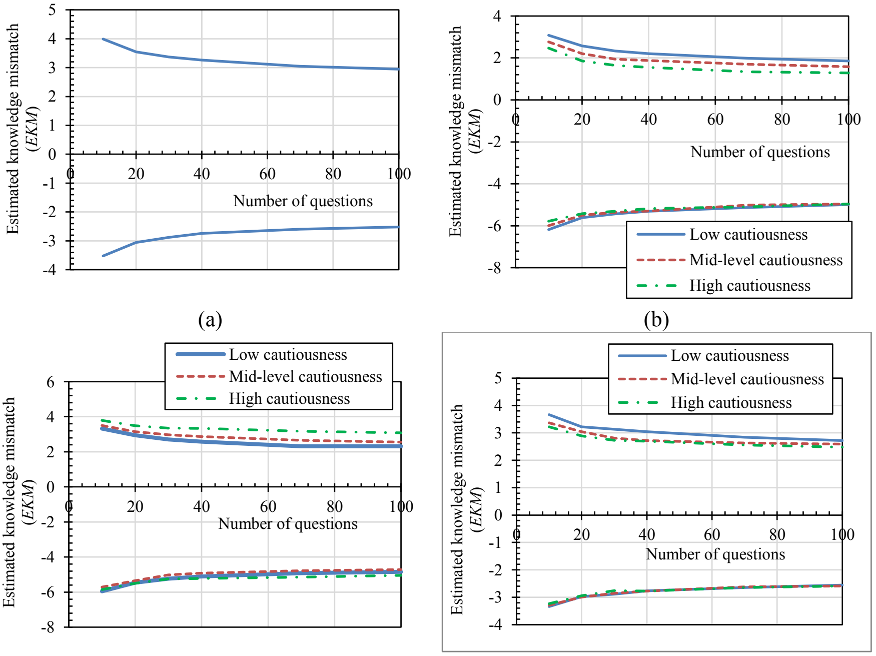
(d1)

(d2)

The increment in the number of questions in the test did not affect the mean value of the *EKM*. Nevertheless, the standard deviation was reduced to an asymptotic value. The Erroneous Knowledge significantly affected the mean values of the *EKM*, reducing them with the increment of the Erroneous Knowledge.

Figure XI(a) shows the envelope for the Number Right scoring method deduced from the results obtained in Figure X. For the cases where the rest of the scoring methods were used, Figures XI(b), XI(c) and XI(d) show their envelopes for the three levels of cautiousness. Negative Marking and Dual Response methods showed a reduction in the upper limit of the envelope when the level of cautiousness of the examinees was increased. The lower limit showed no alteration due to the variation of this input parameter. In the specific case of the Free-Choice method, the upper limit of the envelope showed an increment with the increment of the level of cautiousness.

FIGURE XI. Envelope of the EKM for the (a) Number Right, (b) Negative Marking, (c) Free-Choice and (d) Dual Response scoring methods



(a)

(b)

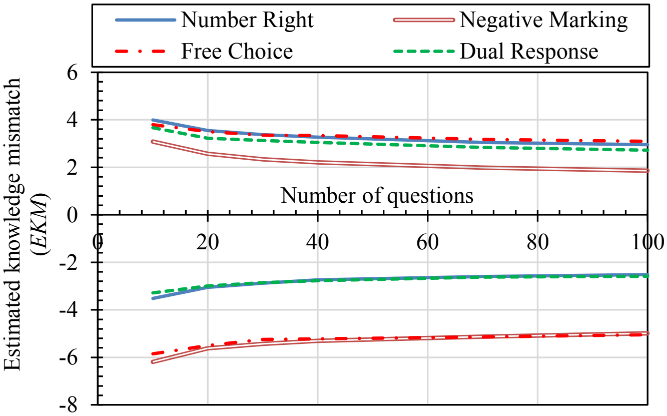
(c)

(d)

Figure XII shows a comparison of the envelopes for all the scoring methods, where the lowest deviation was shown using the Dual Response method. The Negative Marking method showed a high level of penalties with a noteworthy underestimated knowledge followed by the Free-Choice method, which showed the highest deviation in the estimation of the examinees' knowledge.

Received: 29-05-2020 Accepted: 27-11-2020

FIGURE XII. Comparison of the envelopes of the *EKM* for all the scoring methods



Discussion and conclusions

In the previous systematic analysis, the influence of the input parameters in the estimated knowledge obtained using multiple-choice tests was analyzed. An increment in the number of questions reduced the deviation of the knowledge overestimation or underestimation. This is logical from a probabilistic point of view. The Number Right method, generally considered to be a scoring method that overestimates the Real Knowledge of the examinee, shows underestimations of this knowledge when Erroneous Knowledge is strongly present. The Erroneous Knowledge significantly reduced the mean of the *EKM* in all the scoring methods. Thus, this knowledge property considerably affects the reliability of the scoring methods. Specifically, the deviations showed by Negative Marking and Free-Choice methods substantially increase with the presence of Erroneous Knowledge. The level of cautiousness of the examinees influences the upper limit of the envelope of the *EKM* with a maximum influence of one point over 10. Only the Negative Marking method showed lower knowledge overestimation rates (upper limit of the envelope), but at the cost of an extremely critical lower limit of knowledge underestimation.

To analyze in detail the influence of Erroneous Knowledge in the scoring methods, Table 1 shows the mean value of the *EKM* and the standard deviation of this mean value, both out of a maximum knowledge of 10, for all the scoring methods evaluated in this investigation. In the left column of the table, values were obtained by calculating 1000 examinees with input parameters (level of cautiousness, Erroneous Knowledge and Real Knowledge) established at random. In the right column of the table, the same number of examinees were evaluated with randomly selected input parameters but fixing the Erroneous Knowledge to 'none'.

TABLE I. Influence of Erroneous Knowledge in the estimated knowledge and the *EKM*

Scoring method	Random Erroneous Knowledge		Without Erroneous Knowledge	
	μ_{EKM}	σ_{EKM}	μ_{EKM}	σ_{EKM}
Number Right	-0,33	1,35	1,97	0,98
Negative Marking	-2,21	1,77	0,75	0,94
Free Choice	-1,58	2,09	1,76	0,95
Dual Response	-0,41	1,32	1,81	0,83

It can be observed that the mean values of *EKM* were positive for all the scoring methods when no Erroneous Knowledge was established. This means that all scoring methods would overestimate the real knowledge of the examinees. In particular, the overestimated mean value would be from lower than one point out of 10 for the Negative Marking, to nearly two points out of 10 in the case of Number Right. This tendency to overestimate is based on the existence of partial knowledge. This means that even the sanction of Negative Marking is unable to compensate for the points obtained from guessing. In the case of the standard deviation of the *EKM*, no major differences were observed between the scoring methods.

However, when the Erroneous Knowledge is present in the examinees, a significant alteration in the behavior of the scoring methods is seen, showing an underestimation of the Real Knowledge of the examinees. In the specific case of Negative Marking, this underestimation may be higher than two points out of 10. Considering that the presence of Erroneous Knowledge cannot be measured or controlled in an empirical case, the most reliable scoring method would be one that, for a random distribution of the parameters, shows a practically null value of the *EKM* mean and the lowest standard deviation for that coefficient. In this case, the scoring method that best complies with these objectives would be the Number Right method, closely followed by the Dual Response method.

It is important to point out that there are some tests used to certify personnel for industrial or maintenance techniques or internal medicine certification exams, where Erroneous Knowledge must be penalized, because it is more dangerous than Unknown Knowledge. That is why the Negative Marking scoring method would be, without a doubt, the most convenient method in cases that need detection and penalization of the presence of Erroneous Knowledge.

In conclusion, it can be noted that the development of an algorithm for the analysis of the reliability of the scoring methods of multiple-choice tests has provided interesting data about the strengths and weaknesses of each scoring method. The influence of different parameters that are impossible to analyze empirically has been studied using this novel algorithm, opening an interesting research field and showing the potential for using the Monte Carlo method and computer science.

Bibliographic references

Akeroyd, Michael. 1982. “Progress in Multiple Choice Scoring Methods, 1977/81.” *Journal of Further and Higher Education* 6(3):86–90.

Betts, Lucy R., Tracey J. Elder, James Hartley, and M. Trueman. 2009. “Does Correction for Guessing Reduce Students’ Performance on Multiple-Choice Examinations? Yes? No? Sometimes?” *Assessment and Evaluation in Higher Education*.

Budescu, David, and Maya Bar-Hillel. 1993. “To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring.” *Journal of Educational Measurement*.

Burton, Richard F. 2004. “Multiple Choice and True/False Tests: Reliability Measures and Some Implications of Negative Marking.” *Assessment and Evaluation in Higher Education*.

Burton, Richard F. 2005. “Multiple-Choice and True/False Tests: Myths and Misapprehensions.” *Assessment and Evaluation in Higher Education*.

Bush, Martin. 2015. “Reducing the Need for Guesswork in Multiple-Choice Tests.” *Assessment and Evaluation in Higher Education*.

Received: 29-05-2020 Accepted: 27-11-2020

Espinosa, María Paz, and Javier Gardeazabal. 2010. "Optimal Correction for Guessing in Multiple-Choice Tests." *Journal of Mathematical Psychology*.

Hammond, E. J., A. K. McIndoe, A. J. Sansome, and P. M. Spargo. 1998. "Multiple-Choice Examinations: Adopting an Evidence-Based Approach to Exam Technique." *Anaesthesia*.

Hsu, Fu Yuan, Hahn Ming Lee, Tao Hsing Chang, and Yao Ting Sung. 2018. "Automated Estimation of Item Difficulty for Multiple-Choice Tests: An Application of Word Embedding Techniques." *Information Processing and Management*.

Jennings, Sylvia, and Martin Bush. 2006. "A Comparison of Conventional and Liberal (Free-Choice) Multiple-Choice Tests." *Practical Assessment, Research and Evaluation*.

Kurz, Terri Barber. 1999. "A Review of Scoring Algorithms for Multiple-Choice Tests." *Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 21-23, 1999*.

Lin, Chih Kai. 2018. "Effects of Removing Responses With Likely Random Guessing Under Rasch Measurement on a Multiple-Choice Language Proficiency Test." *Language Assessment Quarterly*.

Moon, Jung Aa, Madeleine Keehner, and Irvin R. Katz. 2020. "Test Takers' Response Tendencies in Alternative Item Formats: A Cognitive Science Approach." *Educational Assessment*.

Papenberg, Martin, Birk Diedenhofen, and Jochen Musch. 2019. "An Experimental Validation of Sequential Multiple-Choice Tests." *Journal of Experimental Education*.

Parkes, Jay, and Dawn Zimmaro. 2016. *Learning and Assessing with Multiple-Choice Questions in College Classrooms*.

Riener, Gerhard, and Valentin Wagner. 2017. "Shying Away from Demanding Tasks? Experimental Evidence on Gender Differences in Answering Multiple-Choice Questions." *Economics of Education Review*.

Slepkov, Aaron D., and Alan T. K. Godfrey. 2019. "Partial Credit in Answer-Until-Correct Multiple-Choice Tests Deployed in a Classroom Setting." *Applied Measurement in Education*.

Warwick, Jon, Martin Bush, and Sylvia Jennings. 2010. "Analysis and Evaluation of Liberal (Free-Choice) Multiple-Choice Tests." *Innovation in Teaching and Learning in Information and Computer Sciences* 9(2):1–12.

Contact address: José Calaf Chica. Universidad de Burgos, Escuela Politécnica Superior, Departamento de Ingeniería Civil. EPS Vena, Avenida Cantabria s/n, 09007 Burgos (España). E-mail: calaf@ubu.es

Annex

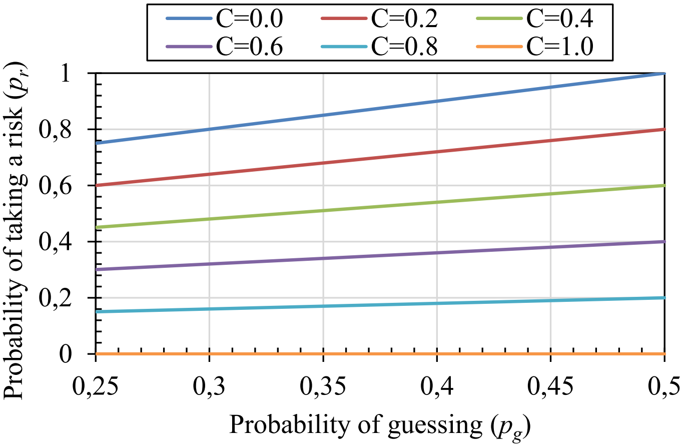
Influence of the level of cautiousness on the probability of taking a risk

The level of cautiousness is the characteristic of an examinee that measures his/her capability to take a risk when he/she has doubts about two or more answer options on a test question. This value, identified by *C*, has a range from 0 to 1. A null value means a bold examinee, and *C=1* means an extremely cautious examinee. As mentioned in the Methodology section, the level of cautiousness controls the probability of taking a risk p_r. This probability p_r controls if the examinee tries to guess the right answer or not. In addition, the examinee could have doubts about two, three or four answer options (in the specific case of a four-choice question). This means that the probability of guessing p_g, if the examinee decides to take a risk, can be lower or higher. This probability of guessing also affects the probability of taking a risk. Thus, p_r is a function dependent on two variables: the level of cautiousness *C* and the probability of guessing p_g. To implement this behavior in the algorithm, the equation (A1) was designed to simulate it. Figure A-I shows this equation (A1). The motivation behind the use of this equation is based on searching for a specific behavior: the more likely it is to guess the right answer (high p_g), the higher the probability of taking a risk (p_r) for all levels of cautiousness. In addition, this equation (A1) showed a reduction of the probability of taking a risk with an increasing level of cautiousness.

$$p_r = (1 - C)\left(0{,}5 + p_g\right) \qquad \text{(A1)}$$

Received: 29-05-2020 Accepted: 27-11-2020

FIGURE A-I. Probability of takinbg a risk vs. the probability of guessing and level of cautiousness (*C*) of the examinee



Analytical probability model for the case for validation

The input parameters for this case for validation were:

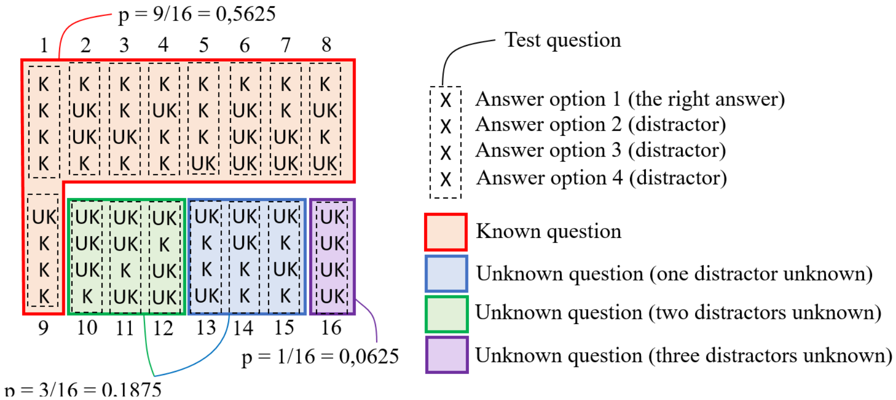
- The number of examinees: 1000.
- Length of the test: 200 questions.
- Real knowledge: fixed to 50% for all the examinees.
- Level of cautiousness: not applicable (a Number Right scoring method was used, so the absence of penalties eliminates any sense of danger).
- Erroneous knowledge: none.
- Scoring method: Number Right.

The four answer options for each question could be only considered as known (K) or unknown (UK), because an erroneously known answer is not possible, because the Erroneous Knowledge has been established previously as ‘none’ (EK = 0). Figure A-II shows the 16 possible scenarios in a four-choice question. Each possible combination for a question is marked with a dashed rectangle, and each answer option is established as known (K) or unknown (UK). In this figure, the first answer option is correct and the other three are the distractors.

Received: 29-05-2020 Accepted: 27-11-2020

The probability of knowing the veracity or falsity of an answer option is equal to 0.5 because the Real Knowledge of the examinees is fixed to 50%. Figure A-II groups the 16 possibilities into three cases: the red group, which gathers together all of the possibilities that are derived from a known question; the blue rectangle, which shows the possibilities with one distractor and the right answer are not known; the green rectangle, which shows the possibilities with two distractors and the right answer are not known; and, finally, the purple rectangle, which shows the single possibility with three distractors and the right answer are not known. Blue, green and purple groups are the questions not known, because the examinee could not deduce without guessing the right answer. The probability of being in green or blue cases would be *p=3/16* for each case. In the purple case, this probability would be *p=1/16,* and finally, the probability of an examinee's knowing the right answer would be *p=9/16*.

FIGURE A-II. Possibilities in a four-choice question (K: known answer; UK: unknown answer)



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
K	K	K	K	K	K	K	K	UK	UK	UK	UK	UK	UK	UK	UK
K	UK	K	UK	K	UK	K	UK	K	UK	UK	K	K	UK	K	UK
K	UK	UK	K	K	UK	UK	K	K	UK	K	UK	K	K	UK	UK
K	K	K	K	UK	UK	UK	UK	K	K	UK	UK	UK	K	K	UK

p = 9/16 = 0,5625
p = 1/16 = 0,0625
p = 3/16 = 0,1875
Test question
X Answer option 1 (the right answer)
X Answer option 2 (distractor)
X Answer option 3 (distractor)
X Answer option 4 (distractor)
🟥 Known question
🟦 Unknown question (one distractor unknown)
🟩 Unknown question (two distractors unknown)
🟪 Unknown question (three distractors unknown)

The questions not known are the cases where guessing could be applied by the examinee. This probability of guessing is equal to p_g=1/2 in the blue cases, p_g=1/3 in the green cases and p_g=1/4 in the purple case.

The probability distribution generally used for several x successes in a sequence of n independent experiments is the binomial probability distribution (BP distribution; see equation (A2)). For example, in a test of

200 questions, the BP for obtaining 100 questions in the blue case would be B(200,100,3/16) = 1.74 x 10^{-23}.

$$B(n,x,p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{A2}$$

If it is assumed that there are 100 blue cases in a test of 200 questions, the probability of correctly guessing 50 questions out of these 100 blue cases would be B(100,50,1/2) = 0.079.

The probability of having 100 questions in the blue case and to correctly guess 50 out of these 100 questions would be the intersection of the previously obtained probabilities: *P* = B(200,100,3/16) x B(100,50,1/2). The analyzed case could be more specific, analyzing all the 200 test questions. For example, the equation (A3) shows the probability of having 30 blue cases with 10 correctly guessed questions, 20 green cases with 7 correctly guessed questions, 10 purple cases with 1 correctly guessed question, with the rest of questions known by the examinee (140 questions):

$$\begin{gathered} P(S=58) = \{B(200{,}30{,}3/16)\cdot B(30{,}10{,}1/2)\times \\ \times\,[B(200{,}20{,}3/16)\cdot B(20{,}7{,}1/3)]\times[B(200{,}10{,}1/16)\cdot B(10{,}1{,}1/4)]\} \end{gathered} \tag{A3}$$

where *S* is the extra score obtained by the examinee. This extra score is calculated as the obtained score in the test (addition of 140 known questions and 10+7+1 correctly guessed questions) minus the score that would represent the Real Knowledge of the examinee (*RK·n* = 0.5 x 200). Thus, *S* = 140+10+7+1–0.5x200 = 58 extra points.

This is not the probability of obtaining 58 extra points, because there are many other probability combinations for obtaining those 58 extra points. Another option could be the probability of having 20 blue cases with 5 correctly guessed questions, 20 green cases with 2 correctly guessed questions, 12 purple cases with 3 correctly guessed questions and the rest of questions known by the examinee (148 questions). Equation (A4) shows the calculation of this probability, where the extra score would also be *S* = 148+5+2+3–0.5x200 = 58 extra points.

Received: 29-05-2020 Accepted: 27-11-2020

$$P(S=58)=\{[B(200{,}20{,}3/16)\cdot B(20{,}5{,}1/2)]\times \\ \times [B(200{,}20{,}3/16)\cdot B(20{,}2{,}1/3)]\times[B(200{,}12{,}1/16)\cdot B(12{,}3{,}1/4)]\} \tag{A4}$$

The sum of all the probability combinations that have S = 58, would give the probability of obtaining an extra score of S = 58 for a test with 200 questions. Equation (A5) shows the calculation of this probability $P(S=58)$.

$$P(S=58)=\sum_{i=0}^{200}\sum_{x=0}^{i}\sum_{j=0}^{200-i}\sum_{y=0}^{j}\sum_{k=0}^{200-j-i}\sum_{z=0}^{k}\{B(200,i,3/16)\cdot B(i,x,1/2)\times \\ \times[B(200,j,3/16)\cdot B(j,y,1/3)]\times[B(200,k,1/16)\cdot B(k,z,1/4)]\} \tag{A5}$$

where *x+y+z+[100-(i+j+k)]* must always be equal to 58.

The most generic case would be the consideration of $S = s$ extra score. The equation (A6) represents this scenario.

$$P(S=s)=\sum_{i=0}^{200}\sum_{x=0}^{i}\sum_{j=0}^{200-i}\sum_{y=0}^{j}\sum_{k=0}^{200-j-i}\sum_{z=0}^{k}\{B(200,i,3/16)\cdot B(i,x,1/2)\times \\ \times[B(200,j,3/16)\cdot B(j,y,1/3)]\times[B(200,k,1/16)\cdot B(k,z,1/4)]\} \tag{A6}$$

where *x+y+z+[100-(i+j+k)]* must always be equal to *s*.

Received: 29-05-2020 Accepted: 27-11-2020