revista de
eDUCACIÓN
Nº 392 ABRIL-JUNIO 2021

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

# Reliability analysis of multiple-choice tests with the Monte Carlo method

José Calaf Chica
María José García Tárrago

GOBIERNO DE ESPAÑA
MINISTERIO DE EDUCACIÓN Y FORMACIÓN PROFESIONAL

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

# Reliability analysis of multiple-choice tests with the Monte Carlo method

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

José Calaf Chica
María José García Tárrago
*Universidad de Burgos*

**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.

**Abstract**

During the twentieth century many investigations have been published about the reliability of the multiple-choice tests for subject evaluation. Specifically, there are a lot of theoretical and empirical studies that compare the different scoring methods applied in tests. A novel algorithm was designed to generate hypothetical examinees with three specific characteristics: real knowledge, level of cautiousness and erroneous knowledge. The first one established the probability to know the veracity or falsity of each answer choice in a multiple-choice test. The cautiousness level showed the probability to answer an unknown question by guessing. Finally, the erroneous knowledge was false knowledge assimilated as true. The test setup needed by the algorithm included the test length, choices per question and the scoring system. The algorithm launched tests to these hypothetical examinees analysing the deviation between the real knowledge and the estimated knowledge (the test score). The most popular test scoring methods (positive marking, negative marking, free-choice tests and the dual response method) were analysed and compared to measure their reliability. In order to validate the algorithm, this was compared with an analytical probabilistic model. This investigation verified that the presence of the erroneous knowledge generates an important alteration in the reliability of the most accepted scoring methods in the educational community (the negative marking method). In view of the impossibility of ascertaining the existence of erroneous knowledge in the examinees using a test, the examiner could penalize its presence with the use of negative marking, or looking for a best fitted estimation of the real knowledge with the positive marking method.

*Keywords:* Test Reliability, Computer Simulation, Scoring, Evaluation, Monte Carlo Methods.

## Introducción

Los test multirrespuesta son ampliamente utilizados en la mayoría de las etapas del sistema educativo de muchos países. Incluso la certificación de competencias y destrezas en multitud de aplicaciones del sector industrial o médico se basan, en numerosas ocasiones, en este sistema de evaluación. Suponen una interesante y útil herramienta cuando el número de alumnos u opositores es en extremo elevado. La fiabilidad del sistema de puntuación utilizado es una cuestión crítica cuando existe una calificación concreta que define la frontera de lo apto para proceder a la certificación o graduación del individuo. Ésta es la esencia que ha motivado la mayoría de los trabajos y publicaciones en torno a los test multirrespuesta (Papenberg, Diedenhofen, and Musch 2019; Parkes and Zimmaro 2016). En éstos se expone una cuestión junto a una selección de múltiples respuestas de las que una de ellas será correcta y el resto distractores. En este sistema de evaluación es de suma importancia el correcto diseño de los distractores (Burton 2005; Hsu et al. 2018), ya que la veracidad o falsedad de éstos debería ser solo clara para un estudiante que conociera la temática evaluada en el test.

Junto a esta clasificación tipológica, existe una amplia lista de métodos alternativos de puntuación de los test. El método más sencillo se conoce como el de "marcado positivo" (Kurz 1999) donde la selección de una opción de respuesta correcta se marca con un punto positivo, mientras que señalar un distractor o no responder ninguna opción no supone puntuación alguna. El principal problema de este sistema de puntuación viene del decalaje generado entre el conocimiento real del estudiante y la estimación de este conocimiento que se extrae del test. Esta sobreestimación del nivel del alumno se debe a la selección de respuestas al azar por parte del estudiante. Al no penalizarse el marcado incorrecto de respuestas, el alumno tiende a marcar al azar las respuestas de aquellas cuestiones que desconoce (Lin 2018). Una forma de corregir esta sobreestimación se alcanza con el uso de un método alternativo de puntuación: el "marcado negativo". Este sistema sanciona la selección de cualquier distractor con una puntuación negativa. Con ello se logra disuadir a los estudiantes de cualquier selección aleatoria de respuestas como forma de alcanzar una calificación superior. Sin embargo, existe otra motivación para seleccionar los distractores: el conocimiento erróneo (Burton 2004). Se entiende por conocimiento erróneo a todo


Recibido: 29-05-2020 Aceptado: 27-11-2020

conocimiento falso que ha sido asimilado por el estudiante como certero. Si se asume que existe un porcentaje de conocimiento erróneo en el saber retenido por el estudiante, parte de la selección de distractores se deberá a este hecho. Con ello el sistema del marcado negativo castigaría por igual la selección de distractores. Además, se observa complejo poder discernir en un estudio empírico la presencia de este tipo de conocimiento en los individuos. Si el marcado negativo castiga por igual ambos comportamientos y la sanción se calibra para eliminar el azar, se deduce que este sistema tenderá a estimar un nivel de conocimiento inferior al que realmente tiene el individuo.

El valor específico de la sanción que se impone por la selección de un distractor en el marcado negativo se establece en base a la teoría de la probabilidad con la finalidad de alcanzar una esperanza matemática nula cuando el test se resuelve por entero al azar (Warwick, Bush, and Jennings 2010). El cálculo de esta sanción se basa en la ecuación (1):

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

donde $p$ es el valor de sanción y $k$ el número de opciones de respuesta por cuestión.

La presencia de otro concepto, el conocimiento parcial (Slepkov and Godfrey 2019), origina una interesante discusión sobre su influencia en los test multirrespuesta. Se basa más en un comportamiento que tienden a tener los individuos que se enfrentan a los test, que a una tipología de conocimiento. Se define como la capacidad de los individuos de discernir algún distractor pero no todos, sin conocer la respuesta correcta ni poder deducirla por descarte (Betts et al. 2009). Pero este comportamiento reduce el número de opciones de respuesta a elegir y, en caso de llegar a un escenario de elección al azar, la probabilidad de acertar se incrementa sensiblemente. Si se tiene en cuenta que la sanción establecida en el sistema de marcado negativo se calcula en base a la ecuación (1) que considera un número $k$ de respuestas posibles, dicha sanción no garantizaría la esperanza matemática nula bajo la presencia del conocimiento parcial. Por lo tanto, la calificación final del test seguiría sobreestimando el conocimiento real del individuo (Budescu and Bar-Hillel 1993).


Recibido: 29-05-2020 Aceptado: 27-11-2020

Existen otros conceptos que influyen en la fiabilidad de los sistemas de puntuación. En concreto, la extensión del test o número de cuestiones planteadas en él. Para poder garantizar la validez de la ecuación (1), la extensión del test debe ser lo suficientemente elevada para que no se incremente la dispersión que puede sufrir el conocimiento estimado frente al real de los individuos evaluados. Hay que tener en cuenta que la ecuación (1) se obtiene de la teoría de la probabilidad y, en consecuencia, es necesario un mínimo número de lanzamientos aleatorios para lograr la esperanza matemática nula que se busca obtener. Pero ese gran número de marcados aleatorios no es en realidad proporcional al número de preguntas o cuestiones incluidas en el test. Un estudiante que tenga amplios conocimientos (elevado conocimiento real) sobre la cuestión tratada en el test tendrá un número limitado de preguntas sin responder y que sean candidatas a ser contestadas por azar. En cambio, ante el mismo test, un individuo sin apenas conocimiento tendrá un banco de preguntas para ser respondidas por azar mucho más amplio. La esperanza matemática nula es más fácil de obtener por individuos de bajo conocimiento real que por estudiantes de elevado conocimiento real. Por tanto, el nivel de conocimiento real del individuo afectaría a la fiabilidad de la ecuación (1) para el caso del marcado negativo.

Hasta ahora se ha planteado que un individuo, en el momento que no sabe responder a una cuestión, intenta acertarla por azar. Pero si se analiza el comportamiento de los individuos ante los test, tal afirmación sería poco acertada. La realidad es más compleja y es aquí donde entra otro parámetro de sumo interés: el nivel de cautela del individuo (Espinosa and Gardeazabal 2010; Riener and Wagner 2017). Hasta ahora se planteaba que la sanción existente en el sistema del marcado negativo lograba anular el uso del azar, pero no hay que olvidar que no todos los individuos se caracterizan por ser igual de arriesgados o cautelosos (Moon, Keehner, and Katz 2020). Además, ese nivel de cautela suele estar adscrito a la personalidad, siendo poco dependiente del conocimiento real (Hammond et al. 1998). En el marcado negativo los individuos muy cautelosos se suelen ver muy influenciados por la existencia de sanciones en los test. La cantidad de preguntas sin contestar tiende a ser elevada frente a unos individuos atrevidos que siguen arriesgándose y optando por completar las cuestiones que desconocen. Con ello, estos últimos acaban sacando provecho de los beneficios que reporta el conocimiento parcial en la respuesta al azar de preguntas. En consecuencia, pueden acabar obteniendo una calificación mayor que la obtenida por un individuo cauteloso del mismo nivel de conocimiento real.

En esta investigación se detallan los cuatro sistemas de puntuación más comunes: los dos anteriormente mencionados, marcado positivo y negativo, y los sistemas de "elección libre" (Jennings and Bush 2006) y de "eliminación" (Bush 2015). El sistema de elección libre destaca por permitir el marcado de más de una respuesta para una misma cuestión. La razón que motiva este sistema es valorar el conocimiento parcial. Por ejemplo, en el caso de una pregunta con cuatro opciones de respuesta, este sistema funcionaría del siguiente modo: si se marca la respuesta correcta, el individuo recibe un punto (3/3); si el individuo marca la respuesta correcta y un distractor, recibe 0,67 puntos ((3-1)/3); si se seleccionan la respuesta correcta y dos distractores, 0,33 puntos ((3-2)/3); si se marcan todas las respuestas, 0 puntos ((3-3)/3); si no se marca la respuesta correcta y únicamente un distractor, se recibe una sanción de -0,33 puntos ((0-1)/3), dos distractores seleccionados, -0,67 puntos ((0-2)/3); y, finalmente, tres distractores seleccionados, -1 punto ((0-3)/3). Existe un método en la literatura, alternativo y semejante al método de la elección libre que se denomina método de la "respuesta doble" (Akeroyd 1982). En él, el sistema de puntuación funciona del siguiente modo: la selección de únicamente la respuesta correcta recibe 1 punto; el marcado de la respuesta correcta y un distractor supone una calificación de 0,5 puntos; y, en caso de seleccionar la respuesta correcta y dos distractores se puntúa con un valor de 0,25. El resto de las combinaciones de marcado no generan puntuación ni sanción alguna. El último sistema de puntuación de test que se mencionó, el de eliminación, se asemeja al de elección libre con la salvedad que el individuo debe marcar las respuestas que cree erróneas (los distractores) en lugar de la respuesta correcta.

Este breve repaso por la multiplicidad de factores que pueden afectar en la fiabilidad de los sistemas de puntuación refleja la complejidad que subyace en cualquier análisis empírico o teórico que se quisiera abordar. Los conocimientos real, erróneo y parcial y el nivel de cautela son parámetros imposibles de valorar en un estudio empírico. La alternativa de los estudios teóricos basados en la teoría de la probabilidad podría tornarse sumamente compleja al intentar tener en cuenta todas las variables aquí descritas. Fue esa la razón por la que en esta investigación se decidió abordar esta cuestión aprovechando las virtudes que ofrecen los algoritmos computacionales. El principal objetivo fue el desarrollo de un código que generara una serie de individuos virtuales caracterizados por una serie de parámetros de entrada (conocimiento real, erróneo y nivel de cautela). Combinando una base de datos de individuos virtuales


Recibido: 29-05-2020 Aceptado: 27-11-2020

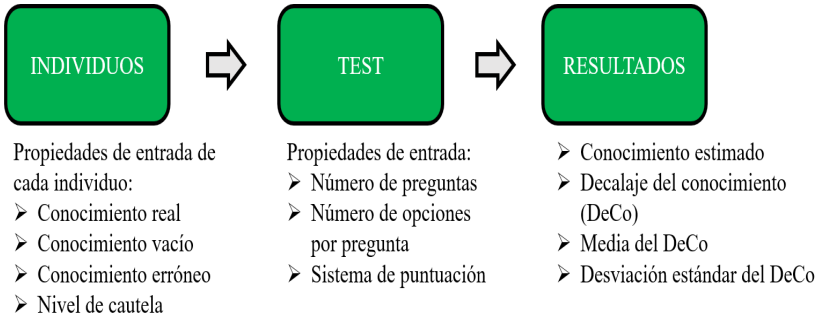
con diferentes diseños de test, el algoritmo ofrecería, como salida, una calificación final o conocimiento estimado. Con ello, el algoritmo permitiría interpretar la influencia de distintos parámetros imposibles de valorar en un estudio empírico.

## Método

El principal objetivo de esta investigación era el desarrollo de un algoritmo que simulara el proceso de marcado de un test multirrespuesta por parte de un individuo virtual. Para la elaboración de este código se hizo uso del lenguaje Python debido a su simplicidad, capacidad y amplia gama de librerías disponibles en módulos intuitivos.

El Grafico I muestra un diagrama básico de flujo del algoritmo, en el que se muestran sus tres bloques principales: los individuos, el test y los resultados. Tanto el bloque de individuos como el del test tienen una serie de parámetros de entrada que definen sus distintas propiedades. El bloque de resultados es el encargado de organizar y suministrar la información de salida obtenida por el algoritmo. Cada uno de los parámetros de entrada y salida es definido en epígrafes posteriores.

**GRÁFICO I.** Diagrama básico de flujo del algoritmo



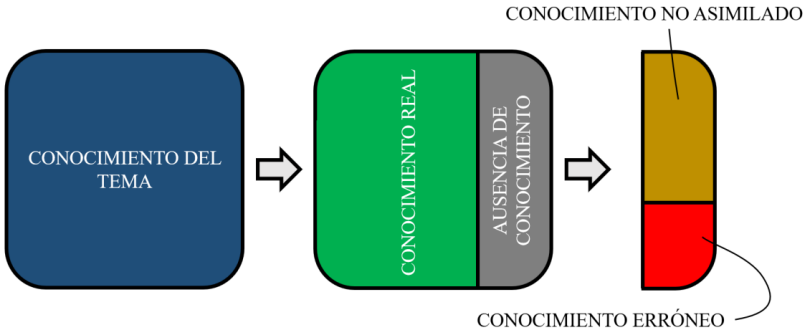

## Propiedades de los individuos

El algoritmo mide un hipotético "conocimiento del tema" que englobaría los saberes del tema evaluado en el test. Los individuos virtuales se


Recibido: 29-05-2020 Aceptado: 27-11-2020

caracterizan por tener asimilado un porcentaje de dicho conocimiento, que es lo que se conoce como "conocimiento real". El Gráfico II muestra una visión esquemática del conocimiento del tema (rectángulo azul) que se clasifica en: conocimiento real (conocimiento asimilado por cada individuo; rectángulo verde); y el resto del conocimiento, que recibe el nombre de "ausencia de conocimiento" (rectángulo gris). A su vez, la ausencia de conocimiento se divide en dos tipologías: el "conocimiento no asimilado" y el "conocimiento erróneo". El conocimiento no asimilado es aquel saber que el individuo no ha logrado retener, y el conocimiento erróneo es el porcentaje de ausencia de conocimiento que ha sido malinterpretado. Estos tres tipos de conocimiento (real, no asimilado y erróneo) son tres parámetros de entrada y son propiedades de cada uno de los individuos virtuales que se generan con el algoritmo. El Gráfico III ilustra esta clasificación de conocimientos en base a un ejemplo práctico: el conocimiento sobre la adición matemática. El conocimiento real se relacionaría con las sumas correctamente realizadas. El conocimiento no asimilado correspondería con aquellas sumas que el alumno no sabe realizar. Finalmente, el conocimiento erróneo correspondería con las sumas incorrectamente calculadas.

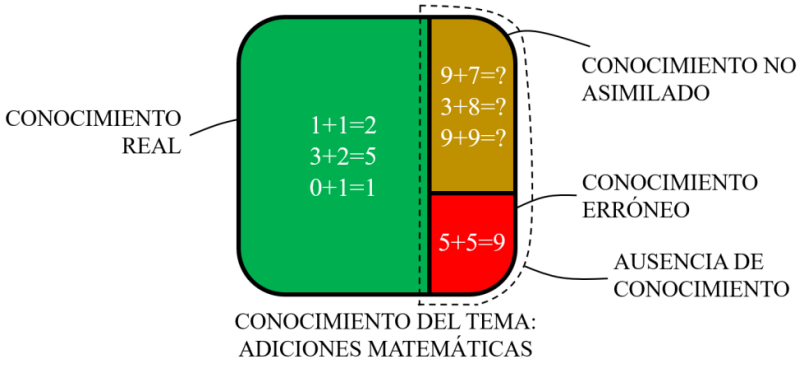
**GRÁFICO II.** Clasificación del conocimiento del tema





Recibido: 29-05-2020 Aceptado: 27-11-2020

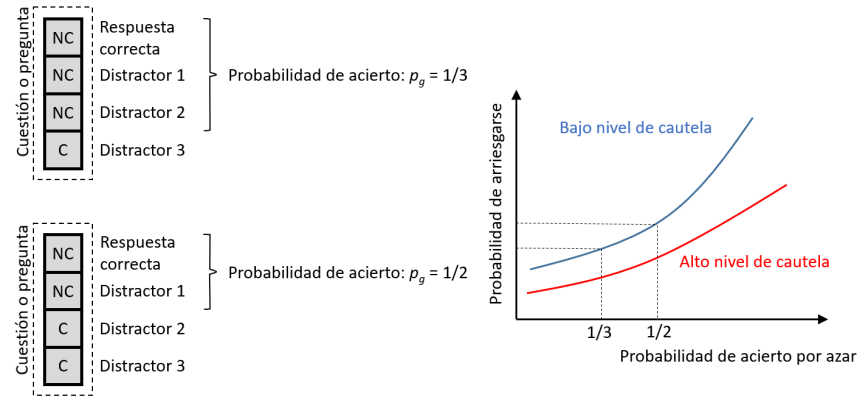
**GRÁFICO III.** Ejemplo de interpretación de la clasificación del conocimiento




Otra propiedad de entrada de los individuos virtuales es su nivel de cautela, que mide la capacidad que éste tiene de arriesgarse a adivinar por azar la respuesta correcta de una cuestión del test. Pero es importante aclarar que la probabilidad de que un individuo decida arriesgarse a responder a una cuestión que desconoce no solo depende de su nivel de cautela, sino también depende de la probabilidad de acierto por azar de cada cuestión. El Gráfico IV muestra un esquema de ejemplo para una cuestión con cuatro opciones de respuesta en el que la primera se identifica con la correcta y el resto con distractores. Para el primer caso expuesto (NC,NC,NC,C), el individuo no conocería la respuesta correcta (identificada con NC), tampoco distinguiría la falsedad o veracidad de dos de los distractores (también identificados con NC), y tan solo conocería un distractor (identificado con C). Con ello, la probabilidad de acierto ante una selección por azar sería igual a $p_g=1/3$. En el caso de que el individuo conociera dos distractores sin discernir el resto de las opciones, la probabilidad de acierto por azar subiría a un valor de $p_g=1/2$. Se deduce que la existencia de conocimiento parcial hace que la probabilidad de acierto por azar sea variable para cuestiones con la misma morfología. Es obvio que, en el caso de tener un individuo con un determinado nivel de cautela, la probabilidad de que se arriesgue a contestar a una pregunta que desconoce será mayor cuanto mayor sea la probabilidad de acierto por azar. Se observa pues esa doble dependencia,

en la que la probabilidad de arriesgarse es dependiente tanto del nivel de cautela del individuo como de la probabilidad de acierto por azar de cada cuestión. El Gráfico IV muestra esa dependencia de dos variables que tiene la probabilidad de arriesgarse. A más probabilidad de acierto, mayor es la probabilidad de arriesgarse y, a menor nivel de cautela del individuo mayor será también la probabilidad de arriesgarse. La metodología específica para la definición de estas curvas de dependencia se encuentra contenida en el Anexo.

**GRÁFICO IV.** Dependencia de la probabilidad de arriesgarse con la probabilidad de acierto por azar y el nivel de cautela




## Propiedades del test

El bloque del test contiene tres parámetros de entrada: la longitud del test, el número de respuestas por cuestión y el sistema de puntuación aplicado. La longitud del test mide el número de preguntas lanzadas a cada individuo. El número de respuestas por cuestión se relaciona de forma directa con el número de distractores que acompañan a la respuesta correcta. Finalmente, el sistema de puntuación (marcado positivo, marcado negativo, elección libre y respuesta doble) es el método con el


Recibido: 29-05-2020 Aceptado: 27-11-2020

que el sistema califica cada pregunta en función del marcado que haya realizado cada individuo. Para el caso específico del marcado negativo y el de elección libre, se requiere un cuarto parámetro relacionado con el nivel de sanción que se asigna a las cuestiones incorrectamente marcadas.

## Salida de resultados

El algoritmo calcula la calificación final del test para cada uno de los individuos y dicho valor se identifica como su conocimiento estimado. Este conocimiento es comparado con el conocimiento real del individuo obteniendo un "decalaje del conocimiento" (*DeCo*) que es igual a la diferencia entre el conocimiento estimado y el conocimiento real (ver ecuación (2)).

$$DeCo = CE - CR \tag{2}$$

donde, *CR* es el conocimiento real y *CE* el conocimiento estimado.

El algoritmo obtiene un conjunto de *DeCo*'s del grupo de individuos analizados y, tras esto, se calcula el valor medio $\mu_{DeCo}$ y su desviación estándar $\sigma_{DeCo}$ (ver ecuación (3)).

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

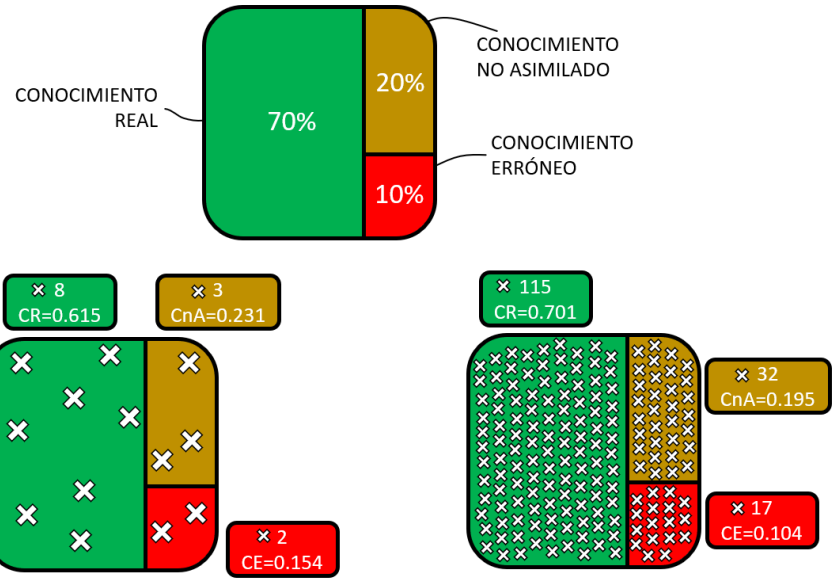
donde *DeCo* es el decalaje del conocimiento y *n* el número de individuos evaluados.

## Simulación de la interacción individuo/preguntas

El algoritmo hace uso del método de Monte Carlo para establecer la interacción entre los individuos y las cuestiones del test. El Gráfico 5 esquematiza cómo se aplica este método estadístico, en el que a

través del uso de una función cuasi-aleatoria se obtienen valores al azar (representados por cruces blancas en la Gráfico V) dentro del espacio contenido por lo que se conoce como "conocimiento del tema". Estos intentos aleatorios pueden caer en las distintas regiones en las que se dividió el conocimiento (real, no asimilado o erróneo). Si el conocimiento real es lo suficientemente amplio es más fácil que el valor aleatorio caiga en la región de este conocimiento. Con ello, el individuo tendería a conocer con más facilidad la veracidad o falsedad de las opciones de respuesta de las cuestiones de un test. En el ejemplo mostrado en el Gráfico V se representa por bloques un individuo con un conocimiento real igual al 70% del conocimiento del tema, un 20% de conocimiento no asimilado y un 10% de conocimiento erróneo. Cada una de las cruces dibujadas representaría el lanzamiento de un valor aleatorio en el algoritmo aquí desarrollado, y representaría la acción del individuo de discernir la veracidad o falsedad de una aserción. En el caso de que el intento cayera dentro del bloque del conocimiento real, el individuo sabría interpretar la certidumbre de la opción de respuesta (marcándose esta opción con una C). Si el intento se ubicara en el bloque del conocimiento no asimilado, el individuo no sabría si dicha opción de respuesta es cierta o falsa etiquetándose ésta con una NC. Finalmente, si el valor aleatorio correspondiera con la región del conocimiento erróneo, el individuo confundiría un distractor como respuesta correcta o viceversa, marcándose esta opción de respuesta con una EC. Se observa en el Gráfico V que cuantos más intentos aleatorios se lanzan, se reduce sensiblemente la diferencia entre la distribución real de porcentajes de conocimientos y los conocimientos estimados por la aplicación del método de Monte Carlo.

**GRÁFICO V.** Método de Monte Carlo aplicado al análisis de una opción de respuesta en un test

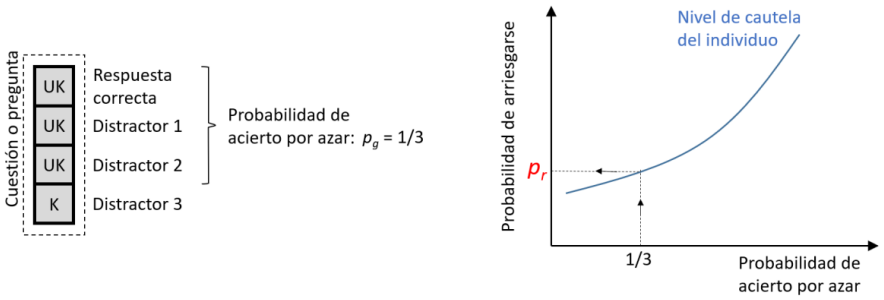



Cuando se lanza la función cuasi-aleatoria para todas las opciones de respuesta de una cuestión de, por ejemplo, cuatro opciones de respuesta, el algoritmo obtiene un identificador del tipo ($A_1$,$A_2$,$A_3$,$A_4$), en el que cada $A_i$ representa a una de las opciones de respuesta ($A_1$ correspondería con la respuesta correcta, mientras que $A_2$ a $A_4$ representarían a los distractores). Cada $A_i$ contiene el resultado de aplicar el método de Monte Carlo indicado con anterioridad con uno de los acrónimos establecidos (C: respuesta conocida, NC: respuesta no conocida y EC: respuesta erróneamente conocida). En el ejemplo (C,NC,NC,NC), la respuesta correcta sería conocida y los tres distractores no conocidos. Con ello, el individuo sería capaz de discernir cual de todas las respuestas sería la correcta y la marcaría. Otro ejemplo de interés sería el caso (NC,C,C,EC). En este caso, el individuo no conocería la veracidad o falsedad de la respuesta correcta, sabría distinguir dos distractores e identificaría erróneamente como cierto al último distractor. Con ello, el individuo marcaría el tercer distractor respondiendo de forma errónea a la pregunta. Finalmente, en el caso de ejemplo (NC,NC,NC,C) en el que únicamente se distingue la


Recibido: 29-05-2020 Aceptado: 27-11-2020

falsedad de un distractor, el individuo podría llevar a cabo un intento de marcado al azar entre las respuestas que le son desconocidas, al no dar por cierta ninguna de las que conoce. Para todas aquellas cuestiones que se encuentran en esta situación, el algoritmo vuelve a utilizar el método de Monte Carlo pero haciendo uso de la probabilidad de arriesgarse del individuo ($p_r$). Como ya se indicó previamente, esta probabilidad $p_r$ es dependiente del nivel de cautela del individuo y de la probabilidad de acierto por azar $p_g$. El primero es una propiedad de cada individuo y el segundo se calcula como $p_g = 1/x$, siendo $x$ el número de opciones de respuesta no conocidas. El Gráfico VI refleja cómo se calcularían tanto la probabilidad de acierto por azar ($p_g$) como la probabilidad de arriesgarse ($p_r$) para el caso específico con identificador (NC,NC,NC,C). La ecuación específica de los distintos niveles de cautela se detalla en el Anexo.

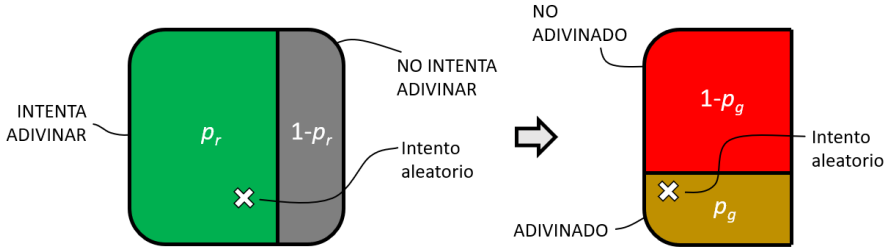
**GRÁFICO VI.** Cálculo de la probabilidad de arriesgarse




El Gráfico VII representa la aplicación esquematizada del método de Monte Carlo para discernir si el individuo decide arriesgarse a adivinar la respuesta correcta de la cuestión. En caso de que el individuo decida intentarlo, el algoritmo vuelve a aplicar el método de Monte Carlo y calcula un nuevo valor aleatorio que es comparado con la probabilidad de acierto por azar $p_g$. En caso de que el intento caiga dentro del espacio ocupado por $p_g$, el individuo marcará la respuesta correcta. En caso contrario, elegirá erróneamente un distractor.


Recibido: 29-05-2020 Aceptado: 27-11-2020

**GRÁFICO VII.** Método de Monte Carlo aplicado para el análisis del proceso de acierto por azar




En una cuestión de cuatro opciones de respuesta por pregunta habría un total de $3^4 = 81$ posibles identificadores que combinaran los casos C, NC y EC. Esto es, una variación con repetición de tres elementos tomados de 4 en 4. El Gráfico VIII muestra el comportamiento que tendrían los individuos para cada uno de los 81 casos posibles. Para todos aquellos en los que el individuo tuviera dudas entre más de una opción de respuesta, se lanzaría el método de Monte Carlo para discernir si el individuo se atrevería a marcar una respuesta por azar y, en caso de ser así, si el individuo lograría acertar o no con la respuesta correcta.


Recibido: 29-05-2020 Aceptado: 27-11-2020

**GRÁFICO VIII.** Clasificación de las 81 posibles combinaciones en una pregunta de cuatro opciones de respuesta



Recibido: 29-05-2020 Aceptado: 27-11-2020

## Proceso de asignación de puntos

El algoritmo utiliza cuatro sistemas de asignación de puntuación: el marcado positivo, el marcado negativo, el de elección libre y el de respuesta doble, cuyos comportamientos han seguido los mismos pasos indicados en el apartado de Introducción. Para llevar a cabo esta asignación de puntos se utilizaron los resultados generados por el proceso de marcado de respuestas previamente comentado. La calificación final de cada individuo virtual se obtiene de la suma de todas las cuestiones correctamente marcadas menos, de las erróneas, la suma de sanciones correspondiente a cada caso.

## Caso de validación y análisis sistemático

El algoritmo fue validado comparándolo con un modelo analítico para un caso particular haciendo uso de la teoría de la probabilidad. Tras esto, el código generado se utilizó para analizar la influencia de cada variable de entrada de individuos y test en el valor medio del decalaje del conocimiento (*DeCo*) y su desviación estándar. En el análisis sistemático se consideraron un conjunto de 720 casos con la siguiente selección de variables de entrada (para el caso estudiado se consideró un test con preguntas de cuatro opciones de respuesta):

- Número de individuos: 1000.
- Longitud del test: 10, 20, 30, 40, 70 y 100 preguntas.
- Conocimiento real: niveles bajo (1), medio (2) y alto (3).
- Nivel de cautela: niveles bajo (1), medio (2) y alto (3).
- Conocimiento erróneo: niveles nulo (1), bajo (2), medio (3) y alto (4).
- Sistema de puntuación: métodos de marcado positivo (NR), marcado negativo (NM), elección libre (FC) y respuesta doble (DR).

Para el conocimiento real los niveles tenían el siguiente significado: el nivel bajo consideraba que el conjunto de individuos tenía un conocimiento real entre el 0% y el 33% del conocimiento del tema; el nivel medio, porcentajes de conocimiento real medio de los individuos entre el 33% y el 66%; y conocimiento real alto entre el 66% y el 100% del conocimiento del tema. Para el conocimiento erróneo, el nivel bajo


Recibido: 29-05-2020 Aceptado: 27-11-2020

significaba que del bloque de "ausencia de conocimiento" (ver Gráfico II) entre el 0% y el 33% estaba formado por conocimiento erróneo, para el nivel medio entre el 33% y el 66% y el nivel alto entre el 66% y el 100%. El nivel nulo significaba que los individuos carecían de conocimiento erróneo dividiéndose el conocimiento del tema únicamente en conocimiento real y conocimiento no asimilado. Para el nivel de cautela, un nivel bajo significaba que los individuos tenían un nivel de cautela *C* entre 0,00 y 0,33, nivel medio entre 0,33 y 0,66 y nivel alto de 0,66 a 1,00. El procedimiento específico seguido por el algoritmo en el uso de este valor *C* para el cálculo de la probabilidad de arriesgarse se encuentra contenido en el apartado Anexo.

Cada uno de los casos analizados en el estudio sistemático se identificó con el ID xx-RKx-Cx-xx-EKx. Como ejemplo, el ID 10-RK1-C1-NR-EK1 representa un test de 10 preguntas, con individuos de bajo conocimiento real, bajo nivel de cautela y nulo conocimiento erróneo, y un sistema de puntuación de marcado positivo.

## Resultados

### Caso de validación

Previo al uso del algoritmo para el análisis de los distintos sistemas de puntuación de los test multirrespuesta, se estudió un caso específico del que fuera posible generar también un modelo analítico basado en la teoría de la probabilidad. Comparando ambos modelos, el analítico y el generado con el algoritmo, pudo validarse el código generado. Las variables de entrada de este caso fueron:

- Número de individuos: 1000.
- Longitud del test: 200 preguntas.
- Conocimiento real: fijo en el 50% para todos los individuos.
- Nivel de cautela: no aplicaba debido a que se utilizó el sistema de puntuación del marcado positivo, donde no existe sanción alguna por marcados erróneos.
- Conocimiento erróneo: nulo.
- Sistema de puntuación: marcado positivo.


Recibido: 29-05-2020 Aceptado: 27-11-2020

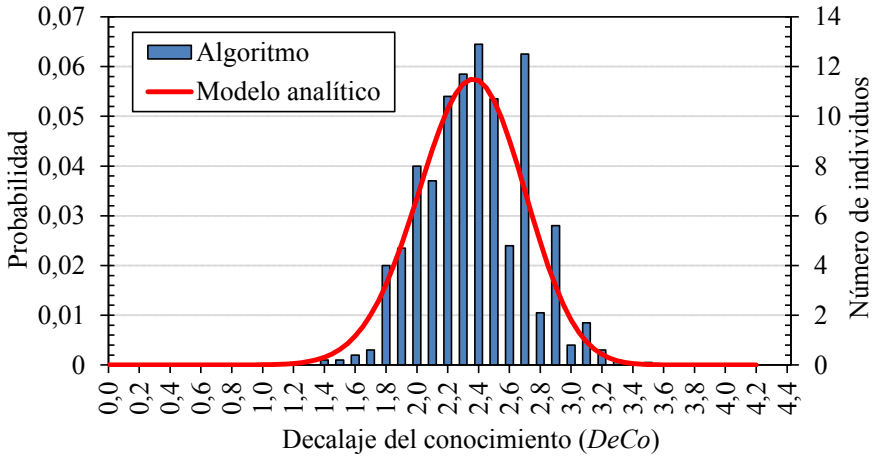
El diseño específico del modelo analítico probabilístico y los pasos seguidos para el desarrollo de la correspondiente ecuación de distribución de probabilidad se han detallado en el Anexo. La ecuación (4) refleja la complejidad del modelo. Hay que tener en cuenta que este primer estudio de validación utilizó el sistema de puntuación más simple (el marcado positivo), eliminando las dificultades que entrañaría la consideración de la cautela o el conocimiento erróneo en el modelo teórico. Queda patente el interés reflejado por el desarrollo de algoritmos basados en el método de Monte Carlo, que permiten analizar escenarios mucho más ricos y complejos, sin tener que abordar y deducir ecuaciones probabilísticas sumamente complejas.

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

donde *x+y+z+[100-(i+j+k)]* debe ser siempre igual a *s*.

El gráfico IX representa la probabilidad de obtener distintos decalajes de conocimiento (*DeCo*) según la ecuación (4) para el caso de estudio especificado más arriba (curva roja). Los decalajes de conocimientos están calculados en base a una puntuación máxima del test en base 10. El diagrama de barras en azul representa el resultado ofrecido por el algoritmo, donde se muestra la distribución de individuos con distintos decalajes de conocimiento para el mismo caso de estudio. La comparación entre ambos modelos, analítico y algoritmo, reflejan el buen funcionamiento del algoritmo como sistema de predicción de calificaciones en individuos virtuales.

**GRÁFICO IX.** Comparación entre el modelo analítico y el resultado ofrecido por el algoritmo

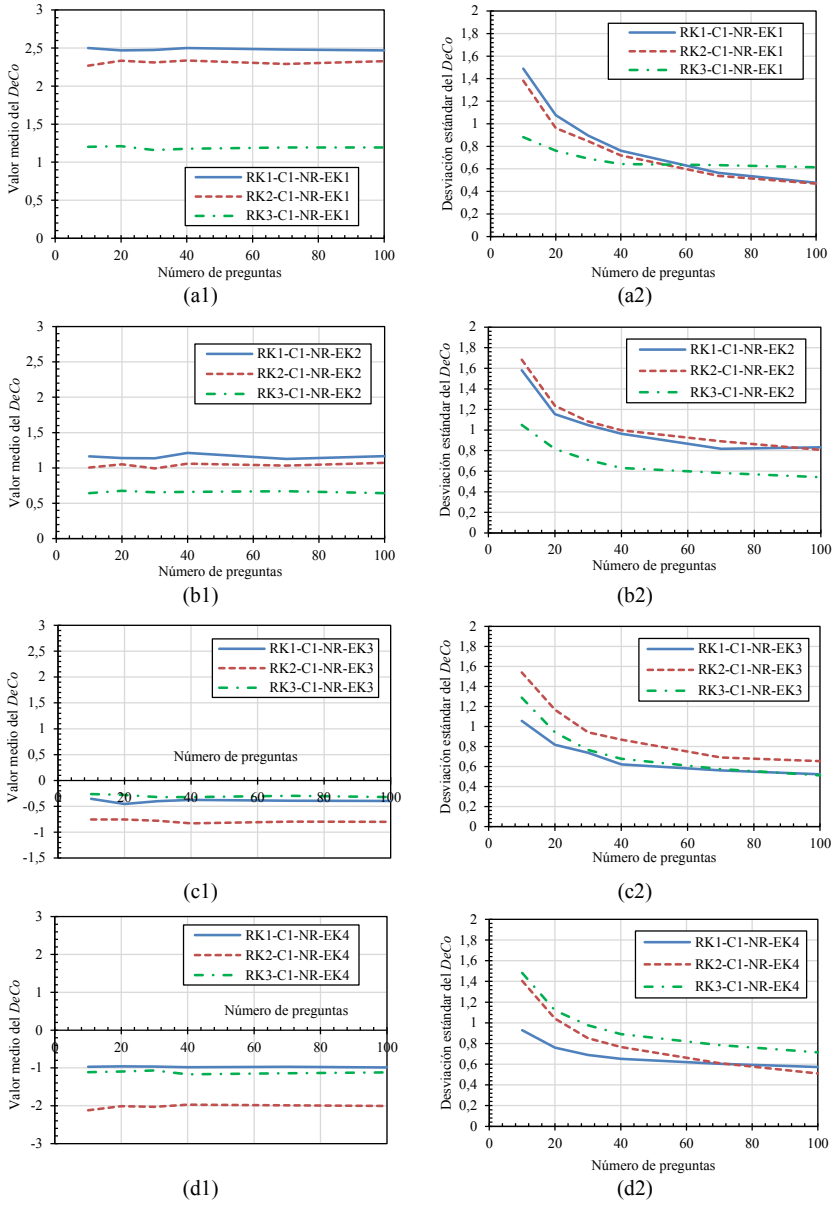



## Análisis sistemático con el algoritmo

Se lanzaron con el algoritmo un total de 720 casos con la finalidad de analizar la influencia de los distintos parámetros de entrada en el conocimiento estimado por el test. Cada uno de los casos reunió a un total de 1000 individuos virtuales calculándose los valores medios y la desviación estándar del decalaje del conocimiento (esto es, la diferencia entre el conocimiento real del individuo con respecto a la calificación obtenida o conocimiento estimado). El Gráfico X muestra a la izquierda los valores medios del *DeCo* y a la derecha las desviaciones estándar del mismo (ambos basados en una calificación máxima de 10 puntos), en función del número de preguntas del test para el sistema de puntuación del marcado positivo. Las líneas continuas azules representan los casos con conocimientos reales bajos (EK1). Las líneas rojas punteadas corresponden con conocimientos reales medios (EK2). Finalmente, las líneas verdes muestran los casos con conocimientos reales altos (EK3). De arriba abajo, cada pareja de gráficos representa un valor de conocimiento erróneo: nulo (EK1), bajo (EK2), medio (EK3) y alto (EK4), respectivamente.


Recibido: 29-05-2020 Aceptado: 27-11-2020

**GRÁFICO X.** Valor medio (1) y desviación estándar (2) del *DeCo* para conocimientos erróneos: (a) nulos, (b) bajos, (c) medios y (d) altos





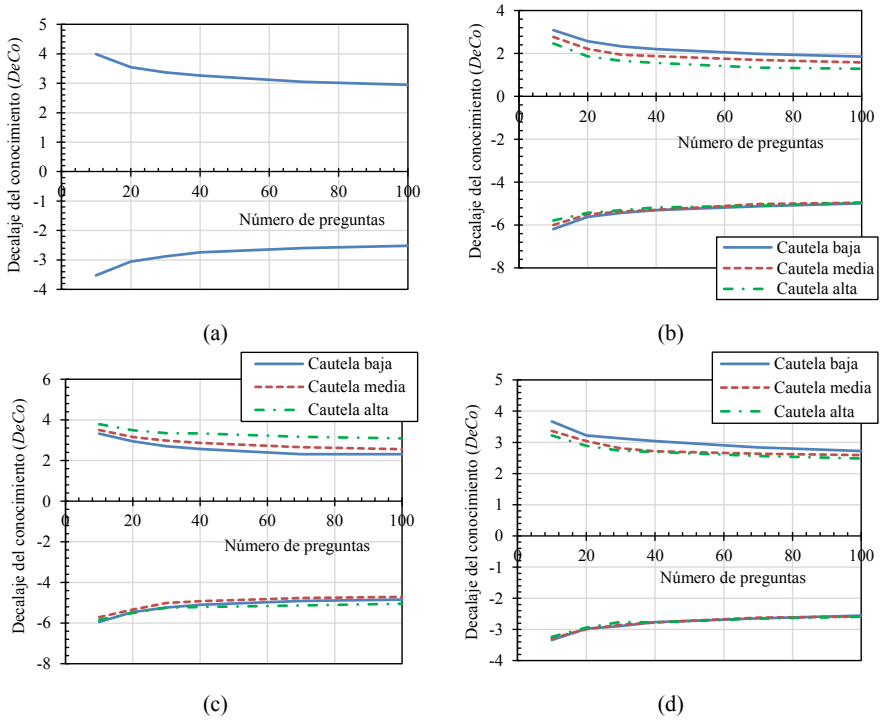
Recibido: 29-05-2020 Aceptado: 27-11-2020

De este primer estudio en relación con el sistema de puntuación del marcado positivo, se observó que el número de preguntas del test no afectaba significativamente al valor medio del *DeCo*, pero sí se reducía sensiblemente el valor de la desviación estándar hasta alcanzar un valor asintótico y estable al aumentar el número de preguntas. La existencia de conocimiento erróneo afectaba de forma visible al valor medio del *DeCo* que llegaba a alcanzar valores negativos para altos niveles de este parámetro. A mayor nivel de conocimiento real de los individuos, el valor medio del *DeCo* tendía a aproximarse a cero.

El Gráfico XI(a) muestra la envolvente que toma el *DeCo* en el caso del sistema de puntuación del marcado positivo. Esta curva se obtuvo de los resultados reflejados en el Gráfico X. Los gráficos XI(b) al (d) muestran las envolventes para el resto de los sistemas de puntuación en donde se tuvo en cuenta la influencia del nivel de cautela de los individuos. Tanto el marcado negativo como el de respuesta doble mostraron una reducción de hasta un punto en el límite superior de la envolvente cuando el nivel de cautela de los individuos se incrementaba. En cambio, el límite inferior no reflejó alteraciones significativas por variaciones en la cautela de los individuos. En el caso específico del sistema de elección libre el comportamiento fue inverso al mostrado por los otros métodos de puntuación: el límite superior de la envolvente aumentaba hasta un punto con el incremento del nivel de cautela.

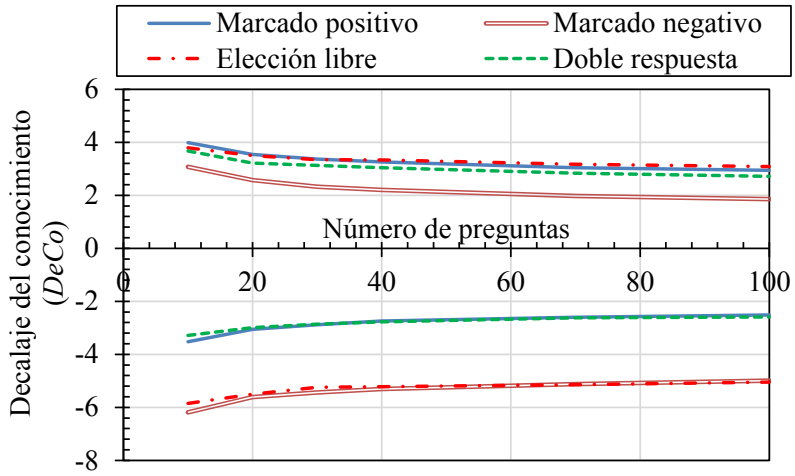

Recibido: 29-05-2020 Aceptado: 27-11-2020

**GRÁFICO XI.** Envolvente del DeCo para los sistemas de puntuación: (a) marcado positivo, (b) marcado negativo, (c) elección libre y (d) respuesta doble




El gráfico XII reúne las envolventes más amplias para cada sistema de puntuación. Se observa que el marcado positivo y el de respuesta doble fueron los métodos que menor dispersión reflejaron. La sanción existente en el marcado negativo producía una destacada subestimación del conocimiento del individuo seguido por el sistema de elección libre que tenía además el mayor nivel de desviación en la estimación del conocimiento.

**GRÁFICO XII.** Comparación de las envolventes del *DeCo* para todos los sistemas de puntuación




## Discusión y conclusiones

En el estudio sistemático previamente realizado se analizó la influencia de cada uno de los parámetros más característicos de un test multirrespuesta en el conocimiento estimado o calificación del test. Se observó que un aumento en el número de preguntas reducía la desviación en la subestimación o sobreestimación del conocimiento. Esto es lógico desde un punto de vista probabilístico. El sistema del marcado positivo, que se ha considerado siempre como un sistema que sobreestimaba el conocimiento real del individuo, puede reflejar subestimaciones del conocimiento en caso de que exista un alto nivel de conocimiento erróneo. La existencia de conocimiento erróneo tiende, en todos los sistemas de puntuación, a minusvalorar el conocimiento real del individuo, así que se trata de una propiedad de los individuos que afecta sensiblemente a la fiabilidad de los sistemas de puntuación. En concreto, los niveles de dispersión de los sistemas de marcado negativo y de elección libre se tornan muy elevados por la presencia de este conocimiento erróneo. En el caso del nivel de cautela, altera la máxima sobreestimación del


Recibido: 29-05-2020 Aceptado: 27-11-2020

conocimiento que cada sistema de puntuación puede llegar a mostrar hasta en un punto sobre 10. Solo el sistema de marcado negativo mostró menores tasas de sobreestimación del conocimiento (límite superior de la envolvente), pero a cambio de un límite inferior de subestimación en extremo elevado.

Con la finalidad de observar con detalle la influencia que tiene en los sistemas de puntuación la presencia de conocimiento erróneo en los individuos, la Tabla 1 muestra el valor medio del *DeCo* y su desviación estándar para una calificación máxima de 10, para los cuatro sistemas de puntuación aplicados a 1000 individuos virtuales con: a la izquierda, parámetros aleatorios del nivel de cautela y de los conocimientos real y erróneo; y a la derecha, parámetros aleatorios del nivel de cautela y del conocimiento real, y nulo conocimiento erróneo.

TABLA I. Influencia del conocimiento erróneo en el conocimiento estimado y el *DeCo*

| Sistema de puntuación | Conocimiento erróneo aleatorio | | Sin conocimiento erróneo | |
|---|---|---|---|---|
| | $\mu_{DeCo}$ | $\sigma_{DeCo}$ | $\mu_{DeCo}$ | $\sigma_{DeCo}$ |
| Marcado positivo | -0,33 | 1,35 | 1,97 | 0,98 |
| Marcado negativo | -2,21 | 1,77 | 0,75 | 0,94 |
| Elección libre | -1,58 | 2,09 | 1,76 | 0,95 |
| Respuesta doble | -0,41 | 1,32 | 1,81 | 0,83 |

Se observa que, al no considerar la existencia de conocimiento erróneo, los valores medios del *DeCo* son positivos para todos los sistemas de puntuación. Esto es, todos los sistemas sobreestimarían el conocimiento real de los alumnos. En concreto, el valor medio de sobreestimación sería menor a un punto sobre 10 en el caso del marcado negativo, llegando a valores de sobreestimación de hasta casi 2 puntos sobre 10 en el caso del marcado positivo. Esta tendencia a la sobreestimación se basa en la existencia del conocimiento parcial, que ayuda a que incluso la sanción del marcado negativo no logre compensar los puntos obtenidos por el marcado por azar. En el caso de la desviación estándar del *DeCo*, no se observaron diferencias significativas de un sistema de puntuación a otro.

Pero en el momento que se considera la presencia de conocimiento erróneo en los individuos, se produce un cambio sustancial del

comportamiento de los sistemas de puntuación llevando a todos ellos al campo de la subestimación del conocimiento real de los individuos. En el caso específico del marcado negativo, esta subestimación puede llegar a ser de más de 2 puntos sobre 10. Teniendo en cuenta que el nivel de conocimiento erróneo de los individuos no puede medirse ni controlarse en un caso empírico, el mejor sistema de puntuación sería aquel que, ante una distribución aleatoria de este parámetro, mostrara el valor más cercano a cero de la media del *DeCo* y la menor desviación estándar de ese coeficiente. En este caso, el sistema que cumple mejor estos objetivos sería el de marcado positivo seguido de cerca del sistema de respuesta doble.

Es importante también aclarar que hay determinados test que tienen una función certificadora para determinadas cualificaciones en aplicaciones industriales o del área de la salud en las que la existencia de conocimiento erróneo podría llegar a ser más peligrosa que la misma ausencia de conocimiento. Es por ello que en caso de que fuera necesario y conveniente detectar y castigar la presencia de conocimiento erróneo en los individuos, de los cuatro sistemas de puntuación el de marcado negativo sería, sin lugar a duda, el más conveniente para ese tipo específico de test habilitantes.

En conclusión, se observa que la creación de un algoritmo para el análisis de fiabilidad de los sistemas de puntuación de los test multirrespuesta ha ofrecido información sumamente interesante sobre las fortalezas y flaquezas de cada sistema. Se ha podido estudiar la influencia de parámetros imposibles de analizar con estudios empíricos, abriendo un interesante campo de estudio y de optimización gracias a la potencialidad del método de Monte Carlo y de la computación aplicados al campo de la educación.

## Referencias bibliográficas

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*.

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.

**Información de contacto:** 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

# Anexo

## Influencia del nivel de cautela en la probabilidad de arriesgarse

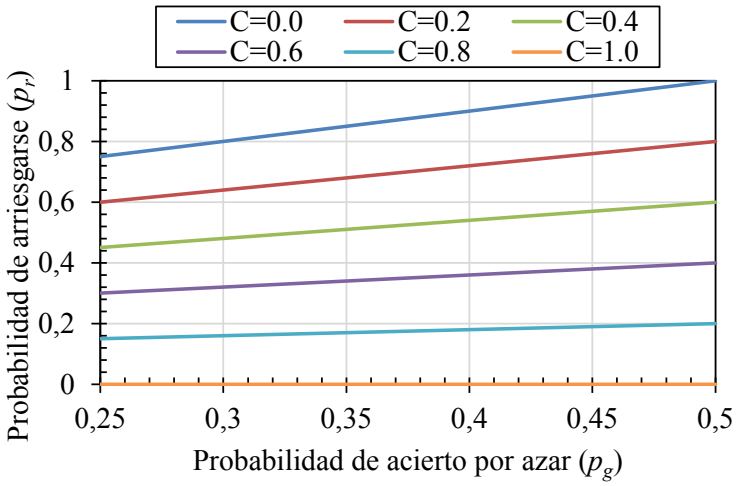
El nivel de cautela es una propiedad del individuo que mide su capacidad de arriesgarse a contestar una pregunta cuando tiene dudas entre más de una opción de respuesta. Esta propiedad, identificada con el acrónimo *C*, muestra valores de 0 a 1. Un valor nulo se identifica con un individuo atrevido y la unidad con un individuo muy cauteloso. Como se menciona en el apartado de metodología, el nivel de cautela modula el valor de la probabilidad de arriesgarse $p_r$. Esta probabilidad controla si el individuo intenta adivinar la opción correcta por azar. Hay que tener en cuenta, además, que el individuo podría tener dudas entre dos, tres o cuatro opciones de respuesta (en el caso de una pregunta de cuatro opciones). Esto significa que la probabilidad de acierto por azar $p_g$ es un valor


Recibido: 29-05-2020 Aceptado: 27-11-2020

variable en el caso de que el individuo intente adivinar la respuesta correcta. Esta probabilidad de acierto por azar también influencia en la probabilidad de arriesgarse. Por lo tanto, $p_r$ es una función dependiente de dos variables: el nivel de cautela $C$ y la probabilidad de acierto por azar $p_g$. Con la finalidad de implementar esta cuestión en el algoritmo se optó por el uso de la siguiente ecuación (A1) que está también ilustrada en el Gráfico A-I. La motivación detrás del uso de esta ecuación se basa en la búsqueda de un comportamiento: que cuanto más probable sea acertar la respuesta correcta (alto $p_g$) la probabilidad de arriesgarse se incremente para todos los niveles de cautela. Además, se buscó que conforme aumentara la cautela, los valores de probabilidad de arriesgarse cayeran consecuentemente. Estas reglas eran por entero cumplidas por la ecuación (A1).

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

**GRÁFICO A-I.** Probabilidad de arriesgarse frente a la probabilidad de acierto por azar y el nivel de cautela



## Modelo analítico de probabilidad para el caso de validación

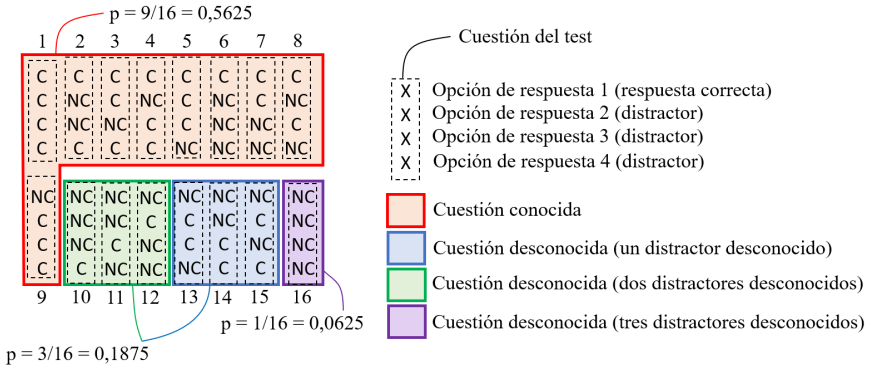
Para este estudio analítico se establecieron los siguientes parámetros de entrada:

- Número de individuos: 1000.
- Longitud del test: 200 preguntas.
- Conocimiento real: fijo en el 50% para todos los individuos.
- Nivel de cautela: no aplicaba al utilizarse el sistema de puntuación del marcado positivo, donde no existe sanción por marcados erróneos.
- Conocimiento erróneo: nulo.
- Sistema de puntuación: marcado positivo.

Las cuatro opciones de respuesta tan solo podían considerarse como conocidas (C) o desconocidas (NC), debido a que, al no existir conocimiento erróneo, ninguna respuesta podía ser erróneamente conocida. El Gráfico A-II muestra los 16 posibles escenarios para una pregunta con cuatro opciones de respuesta. Cada posible combinación se ha encuadrado en un rectángulo punteado, y cada opción de respuesta se ha marcado como conocida (C) o desconocida (NC). De las cuatro respuestas, la superior se considera la respuesta correcta y las otras tres son los distractores. La probabilidad de conocer la veracidad o falsedad de una opción de respuesta era de 0.5 debido a que el conocimiento real de los individuos se estableció y fijó en el 50%. El Gráfico A-II agrupa las 16 posibilidades en tres casos: el grupo rojo, que reúne todas las posibilidades que derivan en que el individuo conoce la respuesta correcta; el cuadro azul, que muestra las posibilidades en las que el individuo conoce dos distractores; el grupo verde, que incluye las posibilidades con un único distractor conocido; y finalmente el púrpura, en donde el individuo no conoce ninguna de las opciones de respuesta. Los grupos azul, verde y púrpura representan las preguntas que el individuo no sabe responder. La probabilidad de encontrarse con un caso verde o azul sería de $p$=3/16 para cada caso. En el caso púrpura, la probabilidad bajaría a $p$=1/16 y, finalmente, la probabilidad de tener un caso en el que el individuo conociera la respuesta correcta sería de $p$=9/16.


Recibido: 29-05-2020 Aceptado: 27-11-2020

**GRÁFICO A-II.** Posibilidades de respuesta en una cuestión con cuatro opciones de respuesta




En los casos en los que el individuo no sabe responder entra en juego la probabilidad de acierto por azar, con valores de $p_g$=1/2 para los casos azules, $p_g$=1/3 para los verdes y $p_g$=1/4 para los púrpuras.

La distribución de probabilidad que se utiliza para el caso de $x$ éxitos en una secuencia de $n$ intentos independientes es la binomial (distribución BP; ver ecuación (A2)). Por ejemplo, en un test de 200 preguntas, el BP para obtener 100 preguntas en el caso azul sería B(200,100,3/16) = 1,74 x $10^{-23}$.

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

Si se asumiera que existen 100 casos azules en un test de 200 preguntas, la probabilidad de acertar por azar 50 preguntas de esos 100 casos azules sería B(100,50,1/2) = 0,079.

La probabilidad de tener 100 preguntas en el caso azul y, a la vez, adivinar por azar 50 preguntas de esas 100 sería igual a la intersección de las dos probabilidades calculadas previamente: $P$ = B(200,100,3/16) x B(100,50,1/2). El caso analizado podría exponerse de forma más concreta, calculando un caso con la totalidad de las 200 preguntas del test. Por ejemplo, la ecuación (A3) muestra la probabilidad de tener 30 casos azules con 10 preguntas correctamente contestadas por azar, 20 casos verdes con 7 preguntas acertadas por azar, 10 púrpuras con 1

pregunta bien respondida por azar y el resto de las preguntas conocidas por el individuo y bien contestadas (140 preguntas).

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

donde *S* es la puntuación extra que obtiene el individuo. Esta puntuación extra se calcula como la puntuación obtenida por el individuo (la suma de las 140 preguntas bien contestadas y las 10+7+1 preguntas acertadas por azar) menos la puntuación que representaría su conocimiento real ($CR \cdot n$ = 0.5 x 200). Por tanto, *S* = 140+10+7+1-0.5x200 = 58 puntos extra.

Pero hay que aclarar que ésta no sería la probabilidad de obtener 58 puntos extra en el test, debido a que existen muchas otras combinaciones de probabilidad con las que obtener esa misma puntuación extra. Por ejemplo, la probabilidad de tener 20 casos azules con 5 respuestas adivinadas, 20 verdes con 2 adivinadas, 12 púrpuras con 3 adivinadas y el resto de las preguntas conocidas por el individuo (148 preguntas). La ecuación (A4) muestra el cálculo de esta probabilidad, donde la puntuación extra sería también de *S* = 148+5+2+3-0.5x200 = 58 puntos extra.

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

La suma de todas las combinaciones de probabilidad que cumplieran *S* = 58, daría la probabilidad de obtener una puntuación extra de *S* = 58 en un test de 200 preguntas. La ecuación (A5) muestra el cálculo de esta probabilidad *P*(*S*=58).

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

donde *x+y+z+[100-(i+j+k)]* debe ser siempre igual a 58.


Recibido: 29-05-2020 Aceptado: 27-11-2020

El caso más genérico sería aquel que considerara que $S=s$ puntos extra. La ecuación (A6) representa este escenario con el que calcular la distribución de probabilidad del caso completo.

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

donde *x+y+z+[100-(i+j+k)]* debe ser siempre igual a *s*.


Recibido: 29-05-2020 Aceptado: 27-11-2020