Desarrollo de un algoritmo en Python para la simulación y análisis de fiabilidad de los test multirrespuesta

Development of a Python algorithm to simulate and analyze the reliability of multiple choice tests to evaluate the student knowledge

María José García Tárrago [1*], José Calaf Chica [2]

[1] Área de Ciencia de los Materiales e Ingeniería Metalúrgica, Departamento de Ingeniería Civil. Universidad de Burgos, España. mjgtarrago@ubu.es*
[2] Área de Ciencia de los Materiales e Ingeniería Metalúrgica, Departamento de Ingeniería Civil. Universidad de Burgos, España. jcalaf@ubu.es

Recibido: 15/01/2020 | Aceptado: 4/04/2020 | Fecha de publicación: 31/08/2020
DOI:**10.20868/abe.2020.2.4461**

RESUMEN

Existe gran número de publicaciones en relación con la fiabilidad de los test multi-respuesta para la evaluación del alumnado en la educación superior. Número de opciones por pregunta, sistemas de puntuación (marcado positivo o negativo), puntuación del conocimiento parcial o cantidad total de preguntas… La combinación de todos estos parámetros es una muestra de la variedad de configuraciones que pueden llegar a establecerse al diseñar un test. ¿Existe algún modelo o configuración óptima? Durante años, los investigadores en innovación educativa han intentado responder a esta cuestión haciendo uso del cálculo de probabilidades y distintas evaluaciones empíricas.

En esta investigación se ha desarrollado un algoritmo basado en código Python con la finalidad de generar una serie de estudiantes hipotéticos con características y habilidades específicas (conocimiento real, nivel de cautela…). Un alto nivel de conocimientos implicaría una alta probabilidad de saber si una de las opciones de respuesta a una cuestión es cierta o no. Un exceso en el nivel de cautela de un alumno estaría relacionado con el nivel de probabilidad que lleva al alumno a arriesgarse a responder a una pregunta de la que no tiene por seguro su respuesta. Ello sería una medida de la capacidad de riesgo del alumno. El algoritmo lanza test a un número específico de alumnos hipotéticos analizando la desviación existente entre el conocimiento real (una característica intrínseca de cada alumno), y el conocimiento estimado por el test.

Una vez desarrollado el algoritmo, se buscó validarlo con el uso de los distintos parámetros de entrada con la finalidad de observar la influencia que estos tenían en la puntuación final del test.

Palabras clave: *test multirrespuesta, Python, evaluación, algoritmo.*

ABSTRACT

There are many literatures related with the reliability of true/false and multiple- choice tests and their application in higher education. Choices per question, positive or negative marking, rewards of partial knowledge or how long they should be… The combination of all these parameters shows the wide set of test setup that each examiner could design. Is there any optimized configuration? An extended educational research has tried to answer these questions using probability calculations and empirical evaluations.

In this investigation, a novel algorithm was designed with Python code to generate hypothetical examinees with specific features (real knowledge, degree of over-cautiousness, fatigue limit…). High knowledge level implies high probability to know whether an answer choice was true or false in a multiple- choice question. Over-cautiousness was related with the probability to answer an unknown question or the risk capacity of the examinee. Finally, fatigue is directly related with the number of questions in the test. Going beyond its upper limit the knowledge level is reduced and the over-cautiousness is increased. The algorithm launched tests to the hypothetical examinees analysing the deviation between the real knowledge (a feature of the examinee), and the estimated knowledge.

This algorithm was used to optimize the different parameters of a test (length of test, choices per question, scoring system…) to reduce the influence of fatigue and over-cautiousness on the final score. An empirical evaluation was performed comparing different test setups to verify and validate the algorithm.

Keywords: *multiple-choice test, Python, evaluation, algorithm.*

1. INTRODUCCIÓN

El método de evaluación de los estudiantes mediante test multi-respuesta ha venido aplicándose de forma amplia y generalizada en la gran mayoría de fases del sistema educativo y, en concreto, es muy común su uso en la educación superior. Durante todo el siglo XX hasta nuestros días, se ha investigado y publicado mucho sobre la fiabilidad de este método de evaluación del conocimiento [1,2]. Los test pueden clasificarse en dos tipologías principales: los test Verdadero/Falso (test VF) y los test multi-respuesta (test MR). En los primeros, se plantea una aserción y el alumno debe indicar la veracidad o falsedad del planteamiento. En los segundos, se plantea parte de una aserción y se incluyen múltiples textos que completan la afirmación, en donde solo uno de ellos es cierto, siendo los demás distractores. Es interesante destacar que cualquier estudio que se lleve a cabo sobre la fiabilidad de un test en su objetivo de cuantificar el conocimiento real de cada alumno, se considera como inexcusable que los distractores estén bien planteados [3]. Esto significa que la falsedad de éstos solo debería ser clara para alguien que conociera la temática planteada en esa cuestión. Si se plantea un distractor cuya falsedad es notoria incluso para alguien ajeno al conocimiento a evaluar, se habría planteado y diseñado el test de forma incorrecta. Con ello, las conclusiones derivadas de una investigación empírica basada en esos test ofrecerían unas conclusiones que podrían ser incorrectas o sesgadas.

Más allá de esta clasificación tipológica, existen distintas formas de puntuar los test una vez cumplimentados por los estudiantes. El más sencillo de todos es el "marcado positivo". En él, una elección correcta se registra y puntúa y el marcado de una opción falsa no supone el aporte de puntuación alguna. Este sistema tiene el problema de generar desviaciones entre el conocimiento real y el conocimiento estimado por el test debido a las respuestas al azar. En concreto, el conocimiento estimado en un test de "marcado positivo" es siempre mayor que el real. La tendencia del alumno es que, una vez ha cumplimentado todas las cuestiones que cree conocer, rellena el resto del test al azar al no suponer la respuesta incorrecta de esas cuestiones ningún tipo de penalización. Una forma de corregir el sesgo de conocimiento real versus estimado que produce este sistema de puntuación, se diseñó el método del "marcado negativo". En él, la selección correcta de respuesta en una cuestión genera una puntuación positiva, al igual que sucede en el "marcado positivo", pero una respuesta incorrectamente seleccionada genera una puntuación negativa. Por tanto, se castigan los errores como sistema con el que evitar la elección de respuestas al azar. Es importante aclarar que siendo el principal objetivo del "marcado negativo" la supresión de ese componente aleatorio que supone la elección al azar de respuestas, este sistema altera o influencia en otro importante aspecto: el conocimiento erróneo [4]. Esto es, el conocimiento que el estudiante ha asimilado como verdadero, pero que es incorrecto. Con ello, parte de las respuestas erróneas de un estudiante pueden, no solo deberse a un marcado al azar para sobrevalorar su puntuación final, sino también venir originadas por una respuesta basada en un conocimiento erróneo. En conclusión, el "marcado negativo" castiga por igual al "conocimiento erróneo" y al "marcado aleatorio" y, por tanto, no puede discernirse ni cuantificarse el peso que cada uno de ellos ha tenido en la puntuación del test. Algunos autores quitan peso al hecho de que

ambas causas raíz se confundan entre sí y sean sancionadas por igual, basándose en el siguiente planteamiento: el conocimiento erróneo es más perjudicial que la falta de conocimiento, ya que hace pensar al alumno que ha adquirido un saber siendo la realidad contraria a tal hecho [4]. Pero es importante puntualizar que, en el caso de las pruebas de evaluación basadas en preguntas de desarrollo, un conocimiento erróneo se puntúa, en general, al mismo nivel que la falta de conocimiento, ofreciendo nula puntuación a ambas conductas. Se observa pues que el sistema de valoración del "marcado negativo" sería más exigente que el establecido en un examen al uso con preguntas de desarrollo, con un sistema de sanciones al "conocimiento erróneo".

El valor específico de la sanción que debería imponerse por cada respuesta incorrecta en un sistema de "marcado negativo" es algo que pocos dudan en establecer en aquel valor que derive en una esperanza matemática nula [5]. Con ello, y suponiendo que cada cuestión correctamente cumplimentada se puntúa con un valor unidad, para los test VF la sanción sería de 1 punto, los test MR de tres opciones de respuesta sería de 0.5 puntos, los de cuatro opciones de 0.33 puntos negativos, etc. El cálculo de estos valores se basa en una sencilla ecuación que deriva de la matemática estadística (ver ecuación 1).

$$p = \frac{1}{k-1} \qquad (1)$$

siendo p el valor de la sanción y k el número de opciones de respuesta que tiene cada cuestión.

Uno de los problemas de este planteamiento se origina cuando se plantea el hecho de que en un test MR, el alumno podría llegar a mostrar un "conocimiento parcial" [6]. Esto es, desconoce la respuesta correcta, pero sí es capaz de discernir alguno de los distractores. Con ello, se reduce el número de opciones de respuesta y, ante una elección al azar, la probabilidad de acierto aumenta sin que la sanción impuesta por el sistema del "marcado negativo" se vea incrementada para compensar dicho evento. En conclusión, el "marcado negativo" es capaz de reducir el componente aleatorio que muestra el sistema del "marcado positivo" pero no por ello queda estadísticamente eliminado. También es importante aclarar que la ecuación (1) aplicada a un test con "marcado negativo" sería válida para reducir la influencia de la selección de respuestas al azar, si el número de preguntas del test es lo suficientemente amplio y, para ser más concretos, el número de preguntas del test que el alumno responde al azar es lo suficientemente amplio. Esto se debe a que el sistema del "marcado negativo" se basa en una esperanza matemática que necesita de un número mínimo de variables (número de preguntas del test) para garantizar que los alumnos no logren alteración alguna en la puntuación final (positiva o negativa) debido al azar en la elección de respuestas. Se introduce por ello otra cuestión a este sistema de puntuación: un estudiante con elevado nivel de conocimientos, que sepa responder a la gran mayoría de cuestiones planteadas en el test, en el caso de que respondiera al azar las pocas preguntas remanentes, podría mostrar una diferencia entre su conocimiento real y el estimado superior al mostrado por un alumno con menor conocimiento real. Esto se debe a que el estudiante con menor conocimiento tendría un banco de preguntas para poder responder al azar muy superior al que tendría el de mayor conocimiento, siendo más fácil para el primer alumno lograr esa esperanza matemática nula, y esa mayor semejanza entre el conocimiento real y el estimado por el test. Si a ello se suma la influencia del "conocimiento parcial", que es más probable que acontezca en el alumno de menor conocimiento, su puntuación final o

conocimiento estimado quedaría menos disperso y ligeramente incrementado.

Hasta ahora, se ha planteado el escenario de un estudiante que respondiera a las cuestiones generadas en el test en su totalidad, donde el hecho de conocer o no la respuesta no fuera una disyuntiva que le llevara a no marcar alguna de las preguntas. Se plantea entonces una cuestión: ¿cuál podría llegar a ser el motivo que llevara a un alumno a no responder a una pregunta? Ante el método del "marcado negativo" lo generaría el temor a la sanción que se impondría ante una respuesta errónea, que es el sentido mismo del método. Pero la esencia con la que se calculó el valor de dicha sanción partía de una premisa: el estudiante respondía a todas las cuestiones, las que conocía la respuesta correcta y las que no. Si el método busca que el alumno altere su comportamiento en base a la amenaza de una sanción y el valor de dicha sanción se basa en que el alumno no altera su comportamiento, se muestra con claridad un problema de planteamiento del método. Este problema deriva en la inclusión de una nueva variable que altera el valor del conocimiento estimado por el test: la cautela del alumno [7]. Ésta es una propiedad inherente de cada alumno y se trata de una variable que nada tiene que ver con el conocimiento real del alumno sobre la temática planteada en el test. Pero su influencia puede llegar a ser notoria en la puntuación final si se aplica el método del "marcado negativo". La amenaza de una sanción afecta mucho más a un alumno cauteloso que a otro más atrevido. El número total de cuestiones marcadas al azar para un mismo nivel de conocimiento real será superior en un alumno que en el otro, generando una desviación entre los conocimientos estimados de esos dos estudiantes. Si se vuelve al caso de estudiantes de bajos conocimientos en el que tenga peso la influencia, antes comentada, del "conocimiento parcial", un estudiante cauteloso nunca marcaría una respuesta, aunque identificara alguno de los distractores de la pregunta. Tan solo marcaría aquellas de las que conociera la respuesta correcta. En cambio, un alumno atrevido tendería con más facilidad a responder al azar y preferentemente aquellas cuestiones en las que existiera "conocimiento parcial". Teniendo en cuenta que en las preguntas respondidas al azar en las que el alumno tiene un "conocimiento parcial" existe una mayor probabilidad de acierto, el alumno atrevido tendería a obtener una sobreestimación de su puntuación final con respecto al alumno cauteloso.

Todas estas cuestiones que reducen o dispersan la fiabilidad del método del "marcado negativo" no eliminan el hecho claro y ampliamente demostrado de que su fiabilidad es superior a la mostrada por el "marcado positivo". Aun así, se muestra la complejidad subyacente del análisis de cada método, sobre todo cuando se llevan a cabo estudios empíricos donde parámetros como el nivel de cautela, el "conocimiento erróneo" o el "conocimiento parcial" tienen influencia en la puntuación final del test y no puede discernirse su existencia o cuantificarse de forma clara cuando se llevan a cabo estudios experimentales. La alternativa de un estudio analítico a través del uso de la matemática estadística podría tornarse complejo si se quisiera tener en cuenta todas las puntualizaciones antes indicadas. Es por ello, que en esta investigación se reflexionó en la posibilidad de utilizar la potencialidad de simulación que un algoritmo puede llegar a ofrecer. Generar un código que creara una serie de alumnos hipotéticos que contuvieran las características y variables de entrada (conocimiento real, nivel de cautela, etc.) que una vez combinadas con un diseño de test, se obtuviera una respuesta en forma de puntuación final o conocimiento estimado. Con ello, podría medirse la fiabilidad de cada diseño de test, sin

que su nivel de complejidad alterara o dificultara en exceso el diseño propio del algoritmo. Este sistema permitiría abordar y analizar cuestiones imposibles de calibrar en un estudio empírico como es la influencia del nivel de cautela en los resultados de un test.

2. METODOLOGÍA

Como ya se comentó en la introducción, el objetivo de esta investigación era el desarrollo de un algoritmo para simular el proceso de realización de un test por parte de un número determinado de alumnos. Python fue el lenguaje elegido para el desarrollo de este algoritmo, debido a su sencillez, capacidad, legibilidad y la amplitud de librerías disponibles actualmente en la red. Para la introducción de datos y la presentación de resultados, se optó por el uso de Excel.

Las variables de entrada establecidas para el algoritmo fueron las siguientes:

➢ *Propiedades de los alumnos:*

- *Número de alumnos.*
- *Conocimiento real. Este conocimiento se midió con el rango [0,1] donde 0 significaba nulo conocimiento y 1 conocimiento completo. Se clasificó este conocimiento real en cuatro posibles niveles: aleatorio, en el que a través de una función aleatoria se establecen al azar niveles de conocimiento para cada uno de los alumnos; bajo, donde la función aleatoria se limitó al rango de conocimiento real [0, 0.33), generando de esta forma alumnos con una horquilla de conocimiento en el primer tercio del rango total; medio, donde se estableció una horquilla para la función aleatoria igual a [0.33, 0.66); y alto, con [0.66, 1].*
- *Nivel de cautela. El nivel de cautela de cada alumno se estableció también con el rango [0,1], donde un valor nulo representaba a un alumno en extremo atrevido y, en el otro límite, el alumno muy cauteloso. Se dispusieron cuatro opciones para esta propiedad bajo el mismo criterio establecido en el conocimiento real, con los niveles aleatorio, bajo, medio y alto, y con las mismas horquillas.*

➢ *Propiedades del test:*

- *Cantidad de preguntas del test.*
- *Número de opciones de respuesta para cada cuestión. Se generaron tres tipos de preguntas: con cuatro, tres y dos respuestas posibles, y se ofreció la posibilidad de combinarlas en un mismo test, pudiendo elegir el número de preguntas de cada tipo.*
- *Sistema de puntuación ("marcado positivo" o "marcado negativo"). En el caso del "marcado negativo", debía también establecerse el valor de sanción para las preguntas contestadas de forma errónea. Ello ofrecía la opción de elegir el nivel de sanción establecido en la ecuación (1) o inferir la consecuencia de establecer otros niveles de sanción distintos.*

El algoritmo iba lanzando el test a cada uno de los alumnos generados. Para cada cuestión formulada se seguía el mismo flujo. A continuación, se muestra el ejemplo para una pregunta con tres opciones de respuesta:

1. Se analiza si el alumno en cuestión conoce o no la veracidad o falsedad de la primera opción de respuesta. Esto se lleva a cabo comparando el valor aleatorio entre 0 y 1 entregado por una función random() y el conocimiento real del alumno. En el caso de que el valor aleatorio sea menor al conocimiento real del alumno, se supondrá que el alumno es capaz de discernir si esa respuesta es real o es un distractor. Tras esto se lleva a cabo el mismo proceso para las otras dos opciones de respuesta.

2. El algoritmo establece a una de las tres opciones como real y las otras dos como distractores. En el caso de que el alumno

conozca la opción real, inmediatamente se le asigna la puntuación de esa pregunta y se pasa a la siguiente cuestión del test. En caso contrario, existen tres posibilidades:

- *Que el alumno conozca todos los distractores. En este caso, aunque no conozca la respuesta real, el alumno la deduciría por descarte. Por ello, se le asignaría automáticamente la puntuación de la pregunta y se pasaría a la siguiente.*
- *Que el alumno conozca un solo distractor.*
- *Que el alumno no conozca ningún distractor.*

3. En el caso de que el alumno se encuentre entre las opciones (b) y (c) se calcula cual sería la probabilidad de fallo (PF) en el caso de que el alumno intentara adivinar la respuesta real al azar entre las opciones que no conoce. Para el caso (b), PF sería igual a 0.5 y para (c) PF = 0.67. En función del número posible de respuestas de cada cuestión, las probabilidades de fallo serían distintas para cada caso.

4. En función del nivel de cautela C del alumno, se calcula una probabilidad de cautela PC, que mide la probabilidad de que el alumno evite responder al azar. Esta se calcula según la ecuación (2).

$$PC = 1 + (C - 1)(1.5 - PF) \qquad (2)$$

Se observa que la probabilidad de cautela es función del nivel de cautela del alumno y de la probabilidad que tiene el alumno de fallar (elegir al azar un distractor). El sentido de esta dependencia se basa en el hecho de que un alumno con determinado nivel de cautela es más probable que evite responder al azar a una pregunta si tiene que elegir entre un mayor número de opciones de respuesta. La Figura 1 muestra la dependencia de la probabilidad de cautela (PC) con el nivel de cautela (C) y la probabilidad de fallo (PF).

5. Se calcula si el alumno finalmente decide arriesgarse a responder al azar a la cuestión. Esto se lleva a cabo comparando la función random() con la probabilidad de cautela (PC) calculada previamente. En el caso que dicha función tenga un valor superior a PC, se deduce que el alumno toma de decisión de arriesgarse. En caso contrario, decide dejar la respuesta en blanco y pasa a la siguiente pregunta del test.

6. En caso de que el alumno haya vencido a su cautela, el algoritmo lanza de nuevo la función random() para compararla con la probabilidad de fallo (PF) de la pregunta. Si random() resulta ser mayor que la probabilidad de fallo (PF), el alumno habrá marcado la respuesta verdadera y se le asignará la puntuación correspondiente a esa pregunta. En caso contrario, dependiendo del sistema de puntuación, sucederá lo siguiente:

- *Para el sistema del "marcado positivo", el alumno no recibirá puntuación alguna ni sanción.*
- *Para el sistema del "marcado negativo", el alumno tendrá una sanción igual a la elegida para el tipo de pregunta.*

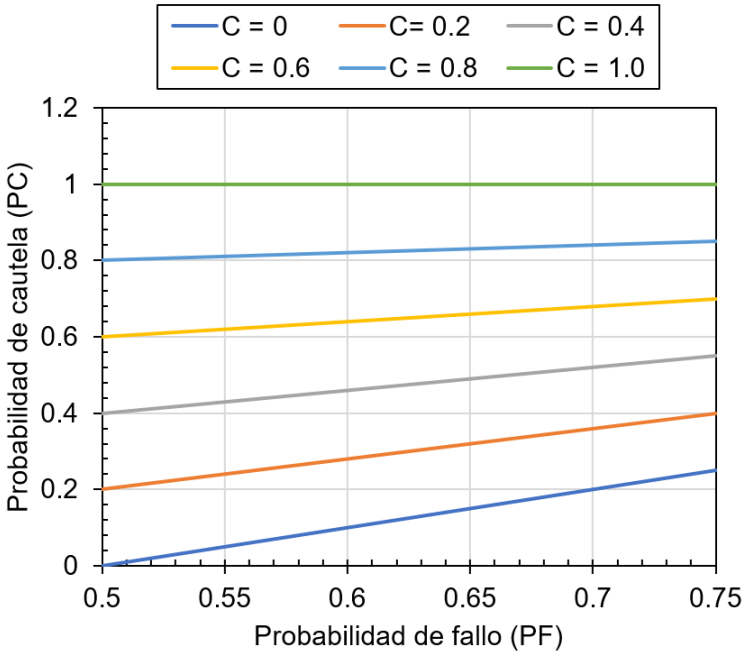


Fig. 1: *Probabilidad de cautela en función de la probabilidad de fallo y la cautela del alumno*

7. Se pasa a la siguiente pregunta del test volviendo al punto 1 de esta descripción. En el momento que se ha pasado por todas las preguntas, se calcula la puntuación final del alumno y se compara con su conocimiento real. Tras esto, se analiza al siguiente estudiante bajo las mismas reglas.

El algoritmo, una vez terminada la evaluación de todos los alumnos hipotéticos, calcula la raíz del error cuadrático medio normalizado (NRMSD) entre los conocimientos reales de cada alumno (CR) y los estimados por el test (CE) (ecuación (3)).

$$NRMSD = 100 \times \sqrt{\frac{\sum_{i=1}^{n}\left[\frac{CR_i - CE_i}{CR_i}\right]^2}{n}} \qquad (3)$$

donde,

CR: conocimiento real del alumno.

CE: conocimiento estimado del alumno por el test (es el valor de la puntuación final normalizada al valor unidad).

n: número total de alumnos evaluados.

También calcula el valor medio de las diferencias entre el conocimiento real y el estimado de los alumnos (ver ecuación (4)), y la desviación de esas diferencias con respecto a su valor medio (ver ecuación (5)).

$$\mu_{DIF} = \frac{\sum_{i=1}^{n}(CE_i - CR_i)}{n} \qquad (4)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}[(CE_i - CR_i) - \mu_{DIF}]^2}{n}} \qquad (5)$$

Las figuras 2 y 3 muestran la ficha de entrada de datos que se crea en Excel y un ejemplo de ficha de salida que se genera una vez lanzado el algoritmo.

Número de alumnos	20	
	Cantidad	Sanción
Preguntas de 4 opciones	0	0.5
Preguntas de 3 opciones	0	0.5
Preguntas de 2 opciones	1000	1

Conocimiento real	2

0: aleatorio
1: bajo
2: medio
3: alto

Sistema de puntuación	1

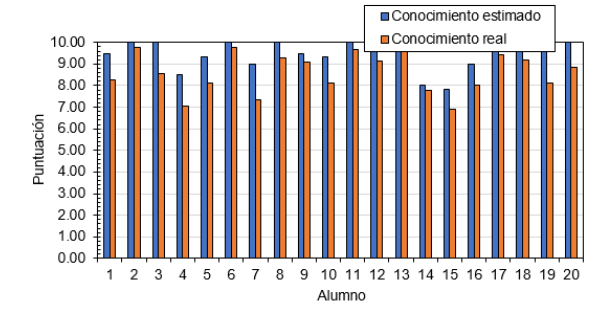
0: Marcado positivo
1: Marcado negativo

Capacidad de riesgo	1

0: aleatorio
1: bajo
2: medio
3: alto

Fig. 2: *Ficha de entrada de datos del algoritmo*

Alumno	Puntos	Nota	CA	CR	RNC	RNCA	NRMSD	Dif. Media	Desv. Dif.
1	19.00	9.50	8.27	0.18	0.00	0.00	108.62	-0.87	0.66
2	20.00	10.00	9.76	0.11	0.00	0.00			
3	20.00	10.00	8.57	0.30	0.00	0.00			
4	17.00	8.50	7.05	0.05	0.00	0.00			
5	18.67	9.33	8.11	0.22	1.00	0.00			
6	20.00	10.00	9.77	0.08	0.00	0.00			
7	18.00	9.00	7.32	0.33	0.00	0.00			
8	20.00	10.00	9.29	0.14	0.00	0.00			
9	19.00	9.50	9.11	0.04	0.00	0.00			
10	18.67	9.33	8.13	0.07	1.00	0.00			
11	20.00	10.00	9.67	0.14	0.00	0.00			
12	20.00	10.00	9.16	0.12	0.00	0.00			
13	20.00	10.00	9.87	0.02	0.00	0.00			
14	16.00	8.00	7.76	0.15	0.00	0.00			
15	15.67	7.83	6.90	0.28	2.00	1.00			
16	18.00	9.00	8.04	0.20	0.00	0.00			
17	20.00	10.00	9.44	0.32	0.00	0.00			
18	20.00	10.00	9.20	0.09	0.00	0.00			
19	20.00	10.00	8.12	0.25	0.00	0.00			
20	20.00	10.00	8.83	0.15	0.00	0.00			
21	17.00	8.50	7.85	0.02	0.00	0.00			
22	20.00	10.00	9.84	0.20	0.00	0.00			
23	15.00	7.50	7.00	0.17	0.00	0.00			
24	17.00	8.50	7.26	0.18	1.00	1.00			
25	19.00	9.50	7.92	0.19	0.00	0.00			
26	17.00	8.50	7.48	0.15	1.00	1.00			
27	18.00	9.00	7.73	0.20	0.00	0.00			
28	19.00	9.50	9.75	0.17	0.00	0.00			
29	15.00	7.50	6.79	0.14	0.00	0.00			
30	18.00	9.00	7.47	0.01	0.00	0.00			
31	20.00	10.00	9.45	0.20	0.00	0.00			
32	19.00	9.50	9.25	0.01	0.00	0.00			
33	15.67	7.83	7.45	0.20	1.00	0.00			
34	15.67	7.83	7.98	0.06	1.00	0.00			
35	19.00	9.50	8.79	0.01	0.00	0.00			
36	19.00	9.50	8.55	0.14	0.00	0.00			
37	20.00	10.00	9.60	0.07	0.00	0.00			
38	16.00	8.00	7.89	0.11	0.00	0.00			
39	13.67	6.83	6.83	0.25	3.00	2.00			
40	16.00	8.00	7.21	0.23	1.00	1.00			
41	19.00	9.50	7.63	0.13	0.00	0.00			



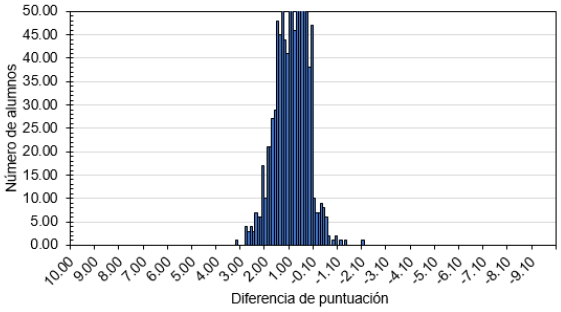


Fig. 3: *Ficha de salida de resultados del algoritmo*

3. RESULTADOS

Una vez generado el algoritmo, se procedió a lanzar una serie de simulaciones con las que poder validarlo. En concreto, se generaron las siguientes pruebas:

- *Tres test con un sistema de puntuación de "marcado positivo" (casos 1 al 3). El objetivo era observar que este formato resulta en un conocimiento estimado superior al real del alumno.*
- *Seis test de "marcado negativo": tres con niveles de cautela bajos (casos 4 al 6) y tres con niveles de cautela elevados (casos 7 al 9). El objetivo era observar que se reducía el conocimiento estimado si lo comparábamos con el "marcado positivo" anterior. También se lanzaron dos tipos de cautela para ver la influencia que ésta tenía en el resultado del test.*

3.1. Análisis del marcado positivo

En el primer estudio (caso 1) de verificación del algoritmo, se lanzaron test con los siguientes datos de entrada:

- Número de alumnos: 1000
- Preguntas con cuatro opciones de respuesta: 20
- Preguntas con tres opciones de respuesta: 0
- Preguntas con dos opciones de respuesta: 0
- Conocimiento real de los alumnos: bajo
- Sistema de puntuación: "marcado positivo"
- Nivel de cautela de los alumnos: no aplicaba al tratarse de un "marcado positivo".

La figura 4 muestra un ejemplo de la comparación entre el conocimiento real y el estimado que se obtuvo para una muestra de 20 alumnos de los 1000 lanzados con el algoritmo. Se observó que el conocimiento estimado era muy superior al real. Esto se debe al efecto producido por las respuestas al azar realizadas por los estudiantes. La figura 5 muestra la distribución de las diferencias entre el conocimiento real y el estimado (nota máxima del test: 10 puntos) y divididas en tramos de 0.1 puntos. Un valor positivo en la diferencia suponía un conocimiento estimado superior al real. En el eje vertical se muestra el número de alumnos incluidos en cada tramo. Se observó que el nivel medio de las diferencias alcanzaba un valor de 2.51 puntos de sobreestimación con una desviación con respecto a esa media de 1.11 y un NRMSD = 273.0%.

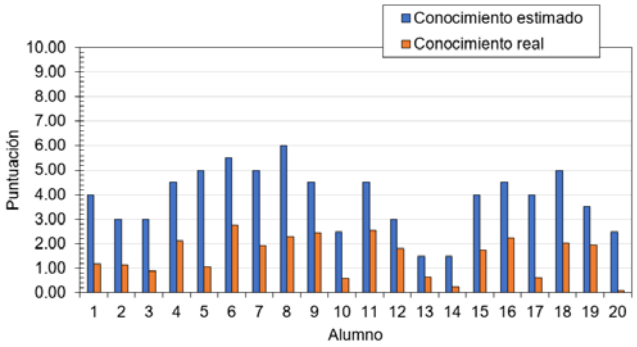


Fig. 4: *Comparación entre conocimiento estimado y real (Caso 1)*

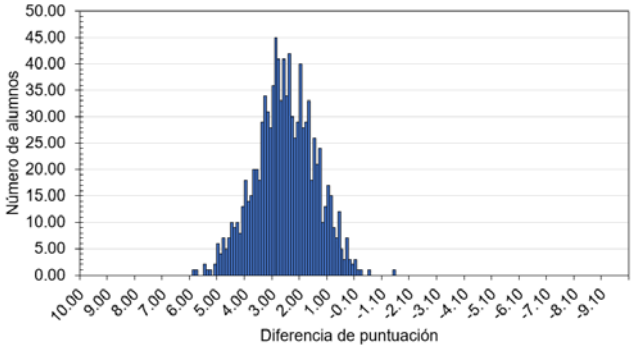


Fig. 5: *Curva de distribución (Caso 1)*

Al aumentar el nivel de conocimiento del alumnado a un valor medio y manteniendo el resto de parámetros de entrada (caso 2), se obtuvieron los resultados incluidos en la figura 6. La media de las diferencias entre conocimiento real y estimado se redujo a 2.32 con una desviación de la media de 1.00 y un NRMSD = 251.3%, menores a los obtenidos en el caso 1.

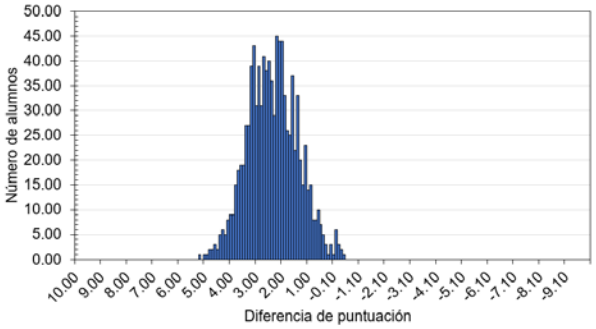


Fig. 6: *Curva de distribución (Caso 2)*

Finalmente, para un grupo de alumnos con conocimientos altos (caso 3) el resultado ofreció la distribución de la figura 7 con un NRMSD = 137.6%. De nuevo la desviación se redujo sensiblemente y la media de las diferencias de conocimiento pasó a un valor de 1.17 de sobreestimación con una desviación con respecto a la mencionada media de 0.73.

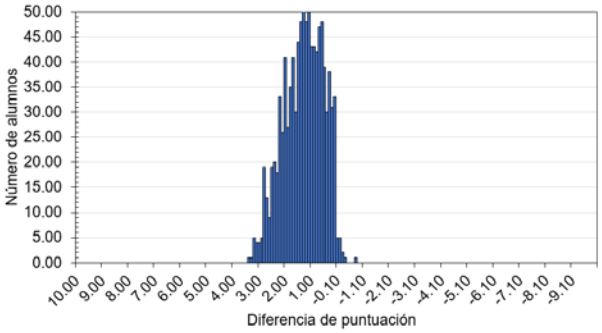


Fig. 7: *Curva de distribución (Caso 3)*

Estos tres casos estudiados con "marcado positivo" mostraron con claridad que este sistema de puntuación sobreestima el conocimiento real de los alumnos. Destaca también la reducción que mostró la diferencia media entre conocimiento real y estimado cuando aumenta el conocimiento real del alumnado. Esto se debe a que el banco de preguntas respondidas al azar disminuye con el aumento de conocimiento. Con ello, se muestra que el "marcado positivo" es más benigno con los alumnos de bajo conocimiento que con los de elevado conocimiento.

3.2. Análisis del marcado negativo

En este caso se estudió y verificó el funcionamiento del "marcado negativo" y se comparó con los resultados obtenidos en los tres casos evaluados en el apartado anterior. Por ello, las variables de entrada fueron las mismas salvo en el sistema de puntuación elegido. Las figuras de más abajo muestran las distribuciones de los casos 4 a 6, en los que se evaluaron alumnos con conocimiento bajos a altos respectivamente. Con respecto al nivel de sanción, se estableció en 0.333 correspondiente a lo calculado según la ecuación (1) para el caso de preguntas de 4 opciones de respuesta, y el nivel de cautela se introdujo inicialmente como bajo.

Caso 4 (conocimientos reales bajos y cautela baja):

- *Media de las diferencias de conocimiento: 0.44 de sobreestimación.*
- *Desviación de las diferencias de conocimiento: 1.35.*
- *NRMSD = 141.91%*

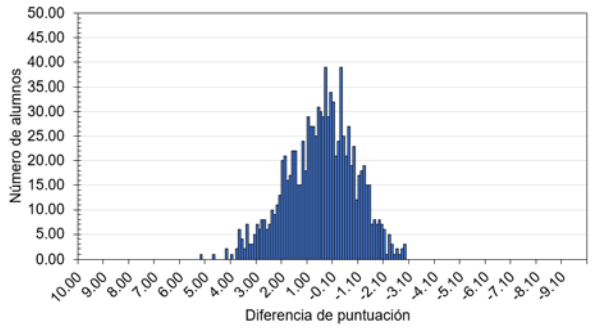
Caso 5 (conocimientos reales medios y cautela baja):

- *Media de las diferencias de conocimiento: 1.20 de sobreestimación.*
- *Desviación de las diferencias de conocimiento: 1.27.*
- *NRMSD = 174.6%*

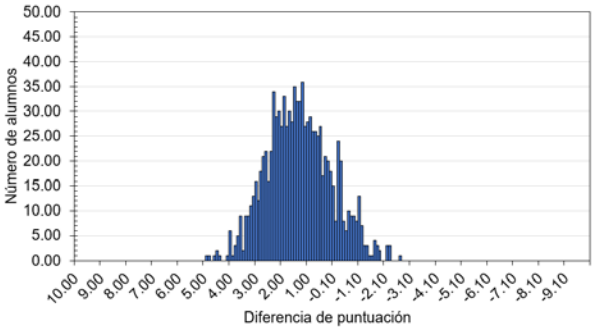
Caso 6 (conocimientos reales altos y cautela baja):

- *Media de las diferencias de conocimiento: 0.98 de sobreestimación.*

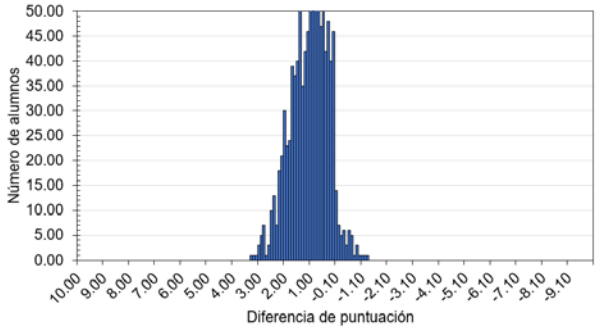
- *Desviación de las diferencias de conocimiento: 0.73.*
- *NRMSD = 122.5%*



(a)



(b)



(c)

Fig. 8: *Curva de distribución: (a) caso 4, (b) caso 5 y (c) caso 6*

El uso del marcado negativo generó una caída en el valor medio de las diferencias entre el conocimiento real y el estimado a cambio de un incremento en la desviación de ese valor medio. Con respecto al NRMSD, el "marcado negativo" logró que se redujera esta desviación, pero fue debido a que ese valor mide la desviación con respecto al conocimiento real y no con respecto al valor medio de la diferencia de conocimientos.

La figura 9 muestra el efecto que tiene el aumento del nivel de cautela al nivel alto (casos 7, 8 y 9). Los valores de media, desviación y NRMSD fueron:

Caso 7 (conocimientos reales bajos y cautela alta):

- *Media de las diferencias de conocimiento: 0.11 de sobreestimación.*
- *Desviación de las diferencias de conocimiento: 0.91.*
- *NRMSD = 91.49%*

Caso 8 (conocimientos reales medios y cautela alta):

- *Media de las diferencias de conocimiento: 0.77 de sobreestimación.*
- *Desviación de las diferencias de conocimiento: 1.17.*
- *NRMSD = 140.1%*

Caso 9 (conocimientos reales altos y cautela alta):

- *Media de las diferencias de conocimiento: 0.87 de sobreestimación.*
- *Desviación de las diferencias de conocimiento: 0.66.*
- *NRMSD = 108.6%*

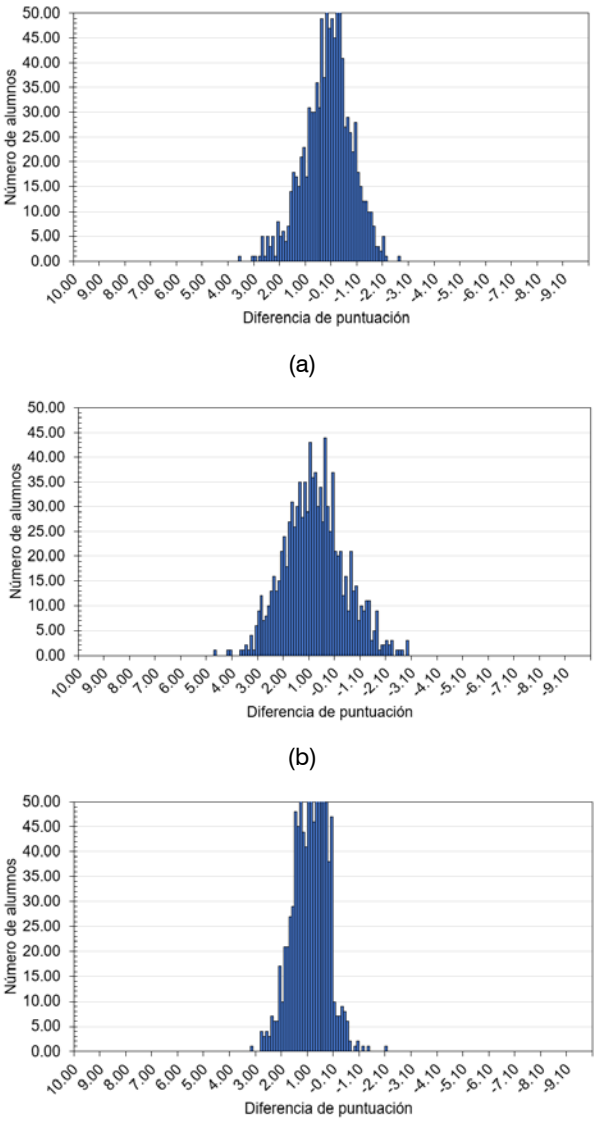


Fig. 9: *Curva de distribución: (a) caso 7, (b) caso 8 y (c) caso 9*

Se observó que el método del "marcado negativo" generaba menores diferencias en la estimación del conocimiento cuando el nivel de cautela de los alumnos era lo suficientemente elevado. La desviación con respecto al valor medio de estas diferencias fue también menor y más ajustada a lo que se obtenía en el caso del "marcado positivo". Se observó que, en el caso de cautela elevada, los alumnos con conocimientos reales bajos eran mucho mejor estimados que los de conocimientos altos, que recibieron un plus de puntuación cercano al 10% del valor máximo.

4. CONCLUSIONES

Se ha diseñado y verificado un algoritmo en Python para el análisis de fiabilidad de los test multi-respuesta. Ello ha permitido analizar la influencia que podría tener el nivel de cautela de un alumno en el conocimiento estimado por un test. El código desarrollado es sensible también a la existencia de conocimiento parcial, y se observan líneas futuras de optimización del algoritmo para que tenga en cuenta aspectos como la fatiga, el conocimiento erróneo u otras tipologías de sistemas de puntuación, como los test multi-respuesta de libre elección [8].

Destaca el hecho de que ambos métodos de puntuación, "marcado positivo" y "marcado negativo", generan de media una sobreestimación del conocimiento. Esto se debe a que el nivel de sanción, que tiende a establecerse en el "marcado negativo" según la ecuación (1), no tiene en cuenta el efecto del conocimiento parcial en la probabilidad de fallo de una pregunta respondida al azar. Ello es lo que genera una tendencia a la sobreestimación de conocimiento, aunque sea menor que en el "marcado positivo".

REFERENCIAS

[1] Ebel, R. L. (1979). Essentials of Educational Measurement. Englewood Cliffs New Jersey, Prentice-Hall.

[2] Lesage, E. & Valcke, M. & Sabbe, E. (2013). Scoring methods for Multiple Choice Assessment in Higher Education – Is it still a Matter of Number Right Scoring or Negative Marking. Studies in Educational Evaluation, vol. 39, pp. 188-193.

[3] Burton, R.F. (2005). Multiple-choice and true/false tests: myths and misapprehensions. Assessment & Evaluation in Higher Education, vol. 30, pp. 65-72.

[4] Burton, R.F. (2004). Multiple-choice and true/false tests: reliability measures and some implications of negative marking. Assessment & Evaluation in Higher Education, vol. 29 pp. 585-595.

[5] Warwick, J., Bush, M. & Jennings, S. (2010). Analysis and Evaluation of Liberal (Free-Choice) Multiple-Choise Tests. Innovation in Teaching and Learning in Information and Computer Sciences, vol. 9 pp. 1-12.

[6] Bush, M. (2001). A Multiple Choice Test that Rewards Partial Knowledge. Journal of Further and Higher Education, vol. 25 pp. 157-163.

[7] Burton, R.F. & Miller, D.J. (1999). Statistical Modelling of multiple-choise and True/False Tests: ways of considering, and of reducing, the uncertainties attributable to guessing. Assessment & Evaluation in Higher Education, vol. 24 pp. 399-411.

[8] Frary, R.B. (1989). Partial-Credit Scoring Methods for Multiple-Choise Tests. Applied Measurement in Education, vol. 2 pp. 79-96.