Revista Iberoamericana de Evaluación Educativa, 2020, 13(1), 225-238.
https://doi.org/10.15366/riee2020.13.1.010

Evaluación de la Validez y Fiabilidad de un Sistema de Test Basado en el Grado de Confianza del Estudiante

Evaluation of the Validity and Reliability of a Test System Based on the Degree of Confidence of the Student

Carlos Valencia Rodríguez *
Yolanda Ortega Latorre
Paloma Huerta Cebrián

Universidad Pontificia Comillas, España

Los exámenes de tipo test basados en la confianza del alumno son escasamente utilizados en el entorno académico español. Este sistema de evaluación hace énfasis tanto en el conocimiento del alumno como en el grado de seguridad que tiene del mismo. Se analizaron los resultados obtenidos en exámenes con el sistema de test basado en la confianza (TBC). El estudio se realizó en una escuela de Enfermería y Fisioterapia durante los cursos comprendidos de 2009 a 2014. Se compararon la validez y fiabilidad del TBC con las obtenidas con el sistema de test con puntuación tradicional por aciertos. Los alumnos adquieren pronto destreza con el nuevo procedimiento de puntuación, escogiendo de forma racional la opción de seguridad más adecuada acorde con su nivel de conocimientos. La validez del sistema de corrección es superior en el TBC respecto al sistema tradicional de test, con una diferencia estadísticamente significativa. La fiabilidad se analiza en cinco grupos de alumnos mediante el coeficiente α de Cronbach, siendo en cuatro de ellos superior la consistencia interna con el TBC e igual en el restante. El test basado en el grado de seguridad ha mostrado ser un procedimiento que aporta ventajas frente al test tradicional. Proporciona mayor información al profesor, los alumnos se adaptan a él rápidamente y ha demostrado mayor validez e igual fiabilidad a las puntuaciones que el sistema tradicional.

Palabras clave: Test basado en la confianza; Evaluación; Validez; Fiabilidad; Aprendizaje; Estudiante universitario; Grado de seguridad.

Test-type exams based on the student's confidence (CBT) are scarcely used in the Spanish academic environment. This evaluation system emphasizes both the student's knowledge and the level of confidences he or she has of that knowledge. The results obtained in CBT were analyzed. The study was conducted in a school of Nursing and Physiotherapy during the academic years covered from 2009 to 2014. The validity and reliability of the CBT were compared to those obtained with a traditional correct-response grading system. Students acquired dexterity with the new scoring procedure quickly, choosing the most appropriate option for confidence level that was in accordance with their level of knowledge in a rational way. The validity of the correction system was higher in the CBT compared to the traditional test system, with a statistically significant difference. Reliability was analyzed in five groups of students using the Cronbach α coefficient, with four of them having higher reliability with the CBT as compared to the traditional grading system and the fifth having the same level of reliability for the two types of tests. Confidence based testing has shown to be a procedure that provides advantages over the traditional test. CBT provides more information to the teacher, students adapt to it quickly and it has greater validity and at least equal reliability to the traditional grading system.

Keywords: Confidence based test; Evaluation; Validity; Reliability; Learning; University student; Security grade.

*Contacto: cvalencia@comillas.edu

issn: 1989-0397
www.rinace.net/riee/
https://revistas.uam.es/riee

Recibido: 23 de marzo de 2019
1ª Evaluación: 19 de septiembre de 2019
2ª Evaluación: 25 de octubre de 2019
Aceptado: 15 de diciembre de 2019

1. Introducción

Las pruebas objetivas de respuestas cerradas de elección múltiple, es decir, los exámenes de tipo test, son cada vez más utilizados en todos los ámbitos donde se intentan evaluar conocimientos o capacidades. Pero los exámenes de tipo test tienen también sus inconvenientes y numerosos detractores. Para corregir las deficiencias de este tipo de exámenes se han introducido numerosos sistemas de test que modifican el procedimiento habitual de puntuación dependiente solo del número de aciertos y fallos. Uno de esos procedimientos es el *test basado en la confianza* (en siglas TBC en castellano y CBT en inglés), procedimiento implantado en la escuela de Enfermería y fisioterapia "San Juan de dios" Universidad Pontificia Comillas desde el año 2009. En este estudio se compara la validez, grado de confianza y fiabilidad del sistema TBC con el sistema convencional de test con puntuación basada en el número de aciertos.

2. Fundamentación teórica

Los exámenes de tipo test son muy utilizados en la evaluación del aprendizaje por su objetividad aparente, por la posibilidad de abarcar un amplio campo de conocimientos y por el menor gasto de tiempo empleado en la corrección, esto último acentuado por la corrección automatizada o por ordenador. Lo sencillo y rápido del proceso, tanto la realización del examen como su corrección, suponen una clara ventaja sobre otros procedimientos de evaluación. En entornos académicos en los que hay un número muy elevado de estudiantes a evaluar, estas ventajas resultan evidentes. Aunque el test de elección múltiple ha sido criticado casi desde el tiempo de su estreno, ha sido especialmente cuestionado durante los años 90 del siglo pasado, cuando un número creciente de educadores, guiados por las teorías constructivistas –propugnadas especialmente por Marton y Säljö (1976a, 1976b), Entwistle, Hanley y Ratcliffe, (1979) y Biggs (1993)– defendieron unos métodos de enseñanza y evaluación que enfatizaban las más altas habilidades de razonamiento, frente a la memorización automática de enunciados. Para sus críticos (Pintrich, 2002), el test de elección múltiple requiere solo identificar la respuesta correcta, pero no requiere un auténtico dominio cognitivo de la materia, a diferencia de lo que ocurre en un examen de preguntas abiertas. El examen de tipo test, para Stanger-Hall (2012), no estimularía las habilidades de pensamiento crítico sino la mera memorización repetitiva.

La objeción más importante que se le hace a este sistema de evaluación es referida a la cuestión de la auténtica validez de las puntuaciones obtenidas por este procedimiento. Otros inconvenientes que se suelen aducir son (Gardner-Medwin, 2006; Morales, 2006; Rippey, 1978):

- La insuficiente discriminación entre los alumnos medianos y los brillantes, con tendencia a homogeneizar las notas.
- El no reconocimiento del conocimiento parcial de la materia evaluada.
- La poca destreza que algunos alumnos bien preparados muestran en el momento de enfrentarse a una elección múltiple.

- La influencia de la personalidad del alumno al optar entre contestar o abstenerse de hacerlo cuando se le da opción de hacer esto último. Aquí la mayor o menos tendencia al riesgo del alumno se convierte en un factor condicionante de la nota.
- La opinión extendida de que para este tipo de exámenes es suficiente un conocimiento superficial de la materia; de hecho, el saber que va a ser sometido a un examen de este tipo podría inducir en el alumno un aprendizaje superficial.

Los test convencionales de elección múltiple son un caso muy especial dentro de una situación más general de toma de decisiones. Al sujeto se le proporciona un cuerpo de información, un conocimiento del campo, un número limitado de opciones para escoger, y este desarrolla un algoritmo de preferencias, haciendo una elección simple entre las opciones de que dispone (Rippey, 1978). Esta decisión, sin embargo, no es muy instructiva sobre el estado de conocimiento del individuo. Una de las cuestiones que se plantean en el sistema de examen de tipo test es la posibilidad de acierto por azar o adivinación. Para neutralizar esa posibilidad, habitualmente se utiliza el procedimiento de penalizar las respuestas erróneas. Este procedimiento anula el efecto del acierto por azar, pero tiene serios inconvenientes. Funcionaría correctamente si el alumno comprendiera perfectamente el enunciado de la pregunta y sólo tuviera dos posibilidades: saber exactamente y con seguridad la respuesta o ignorarla completamente. Pero en la práctica eso no es así; hay diversos grados de conocimiento intermedio entre el conocimiento seguro y el absoluto desconocimiento que revelan grados parciales de dominio de la materia (Morales, 2006; Urosa, 1995). El temperamento del alumno, su situación personal, incluso su género, se convierten en un condicionante de los resultados (Ben-Shakhar y Sinai, 1991; Budescu y Bar-Hillel, 1993; Hassmen y Hunt, 1994). La fórmula habitual de descuento por respuesta incorrecta perjudica a los alumnos menos arriesgados que dejan preguntas sin responder.

En el test convencional la puntuación es la misma al acertar una pregunta, domine el alumno la materia, tenga muchísimas dudas sobre ella o incluso la acierte por azar. El alumno aprende pronto que no hace falta una gran profundidad en el conocimiento para identificar la respuesta correcta. Por ello, los exámenes de tipo test pueden inducir al alumno a desarrollar un aprendizaje superficial (Gardner-Medwin, 2006).

En el examen de preguntas abiertas a desarrollar, el alumno ha de redactar con sus propias palabras la contestación; no puede expresar aquello que ignora; la forma de expresión y exactitud en los conceptos expuestos y el adecuado uso del lenguaje técnico dan al examinador una idea precisa sobre su nivel de dominio de la materia, más allá de que se dé la respuesta correcta a la pregunta. Si el examen es oral, el tono de la voz, la seguridad o dudas del alumno al responder, incluso la expresión corporal, proporcionan pistas adicionales al profesor sobre el nivel de aprendizaje y dominio de la materia por parte del alumno. Por el contrario, la evaluación de los factores mencionados y su transformación en una puntuación adolecen de una notable dosis de subjetividad, a diferencia de la aceptada objetividad del test, que es uno de sus puntos fuertes.

Para evitar los inconvenientes que surgen de la corrección de los test penalizando las respuestas erróneas, se han propuesto varias alternativas. Las más conocidas son las siguientes (Urosa, 1995):

- Número de respuestas correctas

- Selección de alternativas incorrectas
- Técnica de selección del subconjunto
- Responder hasta la correcta
- Fórmula que premia las omisiones
- Fórmula que premia las omisiones y penaliza los errores
- Técnica del grado de seguridad en la selección de la respuesta (confidence based testing)

De acuerdo con la pirámide de aprendizaje de Bloom, el mero conocimiento es solo la base del proceso de aprendizaje; por encima están la comprensión, aplicación, análisis, síntesis y, en lo más alto, la evaluación (Bloom, 1956). Un buen sistema de evaluación debería cubrir, al menos, los dos procesos de la base: el conocimiento y la comprensión, y además, debería ser útil para la autoevaluación del alumno. El sistema de puntuación basado en el grado de seguridad o confianza en la respuesta (en inglés *confidence based test* (*CBT),test basado en la confianza* o TBC en siglas españolas), presenta las siguientes ventajas respecto al tradicional sistema de puntuación basada en el número de aciertos (Gardner, 1969; Gardner-Medwin, 1995, 1998, 2006; Gardner-Medwin y Gahan, 2003; Luetsch y Burrows, 2016):

1. Permite valorar el conocimiento parcial del alumno.
2. Penaliza más los conocimientos erróneos que el simple desconocimiento.
3. Fomenta la sinceridad por parte del alumno.
4. Rebaja sustancialmente la puntuación de los aciertos debidos a conocimientos inseguros o al azar.
5. Fuerza al alumno a reflexionar sobre su falta de confianza en los conceptos.
6. Fomenta un aprendizaje más profundo que los sistemas de examen de tipo test tradicionales.

En el TBC se asignan diversas puntuaciones a la pregunta de tipo test acertada en función de la seguridad que el alumno declara al responder, y también se descuentan, cuando falla, cantidades variables según su grado de seguridad. De esta manera, si responde con mayor seguridad y acierta recibirá una puntuación mayor, pero se arriesgará a una mayor penalización si su respuesta es incorrecta; si declara menor seguridad, obtendrá menos puntuación si acierta, pero también menor (o incluso nulo) descuento si falla.

Este sistema de evaluación hace énfasis tanto en el conocimiento del alumno como en el grado de seguridad y el grado de incertidumbre que tiene en ese mismo conocimiento. La confianza es un componente del proceso de aprendizaje que está siendo objeto de gran interés en los últimos años en el campo educativo, añadiéndose a los constructos de "autoeficacia", "autoconcepto" y "estrés académico", como un predictor importante de los resultados del aprendizaje. Según el modelo de comportamiento basado en el aprendizaje (*Learning-Behaviour Model*), en el cuadro 1 puede observarse la relación entre la cantidad, calidad y seguridad de la información del sujeto con las acciones subsiguientes.

Cuadro 1. Modelo de comportamiento de aprendizaje

SITUACIÓN DEL SUJETO	CONSECUENCIAS
Maestría (conocimiento correcto y autoconfianza)	Acción adecuada
Conocimiento inseguro (conocimiento correcto sin confianza)	Vacilaciones
Desinformación (ausencia consciente de conocimientos)	Parálisis
Mala información (conocimiento incorrecto con autoconfianza)	Acción equivocada

Fuente: Elaboración propia.

La situación más perjudicial es la de la información incorrecta, por debajo del mero desconocimiento; si la desinformación induce parálisis en la acción, que se puede resolver buscando ayuda, la segunda induce acciones equivocadas, lo que es más nocivo (Hunt, 2003). El indicador de confianza añadido a las preguntas del examen TBC, intenta que el alumno aprenda a reconocer y manejar sus dudas. No es extraño, por lo tanto, que los exámenes basados en la confianza se utilicen abundantemente en la enseñanza de las Ciencias de la Salud, sobre todo en países de habla inglesa (Barr y Burke, 2013); si la situación de *acción equivocada* es perjudicial en cualquier campo de actividad humana, lo es especialmente en las Ciencias de la Salud, donde el error puede conllevar consecuencias negativas para el sujeto enfermo.

Se ha comprobado que la confianza es una dimensión robusta que explica diferencias individuales en el rendimiento académico. En estudiantes de secundaria es el mejor predictor de resultados en matemáticas e inglés (Moore y Healy, 2008; Morony, Kleitman, Lee y Stankov, 2013; Stankov, Lee, Luo y Hogan, 2012), física (Sharma y Bewes, 2011) y ciencias (Chang y Cheng, 2008) frente a otros parámetros como autoeficacia docente, ansiedad o autoconcepto. De hecho, está estrechamente relacionada con la autoeficacia y el autoconcepto y es, en realidad, responsable de una gran parte de la varianza predictiva de estas variables, que hasta ahora eran consideradas los mejores predictores de resultados académicos (Morony et al., 2013; Stankov et al., 2012).

En el test basado en la confianza, el alumno podrá maximizar su nota solamente si su conocimiento real es similar a su confianza en su propio conocimiento (Rippey, 1978). Dicho de otra manera, si su metacognición es la adecuada. El fallo en esa parte reguladora del conocimiento, que es la metacognición, llevaría al alumno a cometer errores a la hora de escoger su nivel de seguridad; no es suficiente con saber escoger la respuesta correcta, hay que saber hasta qué punto se conoce la materia evaluada, la densidad y los límites del propio conocimiento. A esta diferencia entre conocimiento y confianza algunos autores la llaman "sesgo de puntuación" o "realismo" y a la situación de que el exceso de confianza sobrepase a los conocimientos reales, "sobreconfianza" (Morony et al., 2013; Stankov, 2000; Stankov, Lee y Paek, 2009).

El sistema de evaluación basado en el grado de seguridad hace posible para el examinador separar la evidencia del conocimiento de la confianza en el conocimiento. Esta separación tiene el potencial de poder proveer mediciones más fiables y válidas del aprendizaje, así como dar información de la tendencia a sobrevaluar o infraevaluar el conocimiento de cada uno de los alumnos (Rippey, 1978). Algunos autores (Ahlgren, 1970), han encontrado evidencias de que el conocimiento que es señalado por el alumno como "seguro" en este tipo de exámenes permanece más tiempo en la memoria que el inseguro, con lo que los

procedimientos de puntuación basados en la seguridad darían también información sobre el aprendizaje a largo plazo de alumno. El test basado en el grado de seguridad proporciona al docente un método de evaluación más potente y completo que el sistema habitual de test basado solo en el número de aciertos.

El primer examen con puntuación basada en el grado de seguridad publicado lo realizó Kate Hevner, investigadora en el campo de la psicología cognitiva musical en Estados Unidos (Hevner, 1932). En ese caso se trataba de un examen de test de tipo "verdadero/falso". El procedimiento no vuelve a utilizarse hasta más de 30 años después (Ebel, 1968; Gardner, 1969; Khan, Davies y Gupta, 2001). En tiempos más recientes, ha sido el grupo de Gardner-Medwin (1995, 1998, 2006), en la escuela de medicina del University College of London, quien ha estudiado más profundamente este sistema. En este trabajo se seguirá parcialmente, su método, muy utilizado en el Reino Unido en carreras relacionadas con las Ciencias de la Salud (Barr y Burke, 2013; Gardner-Medwin y Curtin, 2007; Gardner-Medwin y Gahan, 2003).

A continuación (cuadro 2) puede observarse la puntuación que utiliza Gardner-Medwin para test con tres o más opciones de respuesta:

Cuadro 2. Puntuación del TBC con 3 opciones de respuesta

NIVELES DE SEGURIDAD	ALTO	MEDIO	BAJO
Respuesta correcta	+ 3	+ 2	+ 1
Respuesta errónea	- 4	- 1	0

Fuente: Elaboración propia.

Como se puede comprobar, la máxima penalización se produce cuando el alumno da una respuesta errónea con total seguridad. Cuando el alumno reconoce no estar seguro, no recibe penalización en la respuesta incorrecta.

Los estudios han mostrado la validez y fiabilidad de estos métodos, aunque no de forma unánime. Según algunos autores (Gardner, 1969; Hambleton, Roberts y Traub, 1970; Kansup y Hakstian, 1975), la correlación de los resultados obtenidos en las pruebas con valoración del nivel de seguridad con los obtenidos con otros sistemas de evaluación (comprobación habitual de la validez) es mejor que cuando se puntúa simplemente mediante aciertos. Para otros (Ahlgren, 1970; Gardner-Medwin, 2006), la consistencia interna (o fiabilidad) de los resultados aumenta también. Otros estudios, en cambio, no han mostrado incremento ni en la validez (Rippey, 1978) ni en la fiabilidad de los resultados del examen con este procedimiento (Frary, 1982).

Este sistema de test, con corrección basada en la seguridad, no ha sido prácticamente utilizado en nuestro país. En el presente trabajo se presentan los resultados obtenidos con él, se analiza si los alumnos se adaptan bien al nuevo sistema, y se determina su validez y fiabilidad como método de evaluación en comparación con el sistema de puntuación habitual (test basado sólo en el número de aciertos).

3. Métodos

El trabajo se realiza entre 2009 y 2014 utilizando los exámenes de los alumnos de la Escuela Universitaria de Enfermería y Fisioterapia San Juan de Dios de la Universidad Pontificia Comillas. Los exámenes corresponden a las asignaturas "Afecciones Médico-

Quirúrgicas I y II" de Fisioterapia y "Enfermería Médico-Quirúrgica II" y "Fisiopatología General" de Enfermería.

Los objetivos del presente estudio son:

1. Comprobar si los alumnos escogen adecuadamente el nivel de confianza en sus respuestas, es decir, de forma proporcional a sus conocimientos.
2. Comprobar la fiabilidad de TBC comparándola con la del sistema tradicional de corrección.
3. Comprobar la validez de TBC comparándola con la del sistema tradicional de corrección.

El test, formado siempre por preguntas con tres opciones de respuesta y de ellas una sola válida, se corrige según el sistema de puntuación basado en el grado de confianza. En la presente investigación se sigue la tabla de puntuación de Gardner-Medwin (1995) (tabla 2) con una pequeña modificación, pues se reduce la penalización por respuesta incorrecta en la opción de máxima seguridad, ya que se considera que una excesiva penalización de las respuestas incorrectas podría generar una tendencia hacia la excesiva prudencia por parte del estudiante (cuadro 3).

Cuadro 3. Puntuación utilizada en este trabajo con 3 opciones de respuesta

NIVELES DE SEGURIDAD	ALTO	MEDIO	BAJO
Respuesta correcta	+ 3	+ 2	+ 1
Respuesta errónea	−3	−1	0

Fuente: Elaboración propia.

Los términos elegidos para indicar el grado de confianza fueron: "totalmente seguro" para el nivel máximo, "bastante seguro" para el intermedio y "no seguro" para el mínimo. El alumno escoge la opción de respuesta y luego escoge su nivel de seguridad. El formato de cada pregunta se muestra en la figura 1.

1.-Enunciado de la pregunta

A) Respuesta primera.

B) Respuesta segunda.

C) Respuesta tercera. RESPUESTA CORRECTA: ……

Totalmente seguro **Bastante seguro** **No estoy seguro**

□ □ □

Figura 1. Formato de cada pregunta del test
Fuente: Elaboración propia.

También se corrigen los mismos exámenes según el sistema tradicional, por aciertos con penalización por respuesta incorrecta sin tener en cuenta el grado de confianza, según la fórmula convencional para preguntas con tres opciones de respuesta (cuadro 4).

Cuadro 4. Procedimiento de puntuación convencional

PUNTUACIÓN POR ACIERTOS	PUNTUACIÓN
Respuesta correcta	+1
Respuesta incorrecta	-0'5

Fuente: Elaboración propia.

Por tanto, en cada pregunta de cada examen se registran dos notas, la obtenida mediante el TBC y la obtenida por el sistema convencional de corrección de test.

Para estudiar si el alumno escoge su nivel de confianza con criterios racionales se calcula el porcentaje de respuestas correctas en cada uno de los tres niveles de seguridad. Este cálculo se hace sobre los exámenes de 251 alumnos en 5 exámenes que se hicieron entre 2009, 2013 y 2014. La elección del nivel de confianza se considera adecuada si el porcentaje de respuestas correctas era superior al 80% de las marcadas con máximo nivel de seguridad, estaba entre el 60% y el 80% en las marcadas con nivel intermedio de seguridad y era inferior al 60% en las preguntas señaladas con el nivel mínimo de confianza.

Para la fiabilidad se calcula el coeficiente α de Cronbach en los test corregidos tanto por el sistema basado en el grado de confianza como por el sistema tradicional. Para calcular este coeficiente se utilizaron los mismos exámenes que para el cálculo de la adecuación en la elección del nivel de confianza, 251 sujetos en 5 exámenes realizados entre 2009 y 2014. Para los análisis de los estadísticos básicos y de los índices de correlación se utiliza la tabla de herramientas Excel. El coeficiente α de Cronbach se calcula aplicando la fórmula correspondiente a los datos de las varianzas proporcionados por la tabla de herramientas Excel. Para calcular la significación estadística de la diferencia entre los coeficientes de correlación se utiliza el programa *SISA online statistical analisis*, disponible en Internet.

Para la determinación de la validez se utilizan 215 exámenes de los cursos comprendidos entre 2009 y 2014. En ellos se comparan los resultados obtenidos en el test según los dos procedimientos de corrección –según confianza y según el número de aciertos- con una variable-criterio que es la nota obtenida por cada alumno en la segunda parte del examen, que era un examen de preguntas abiertas. Se utiliza como variable-criterio el examen de preguntas abiertas por ser el procedimiento más antiguo y probado de puntuación. Los alumnos realizaron un examen mixto formado por de 30 preguntas de tipo test y 3 preguntas abiertas de mediana extensión, ambas partes sobre el mismo contenido a evaluar.

En los cinco exámenes estudiados se recogen en tablas Excel los siguientes datos de cada sujeto:

1. Puntuación obtenida por cada alumno con el TBC
2. Puntuación obtenida por cada alumno con el test corregido según el número de aciertos
3. Puntuación obtenida por cada alumno en la parte del examen de preguntas abiertas.

Como criterio de *validez* se establece el obtener una correlación superior a 0,5 y significativa entre la nota obtenida en el test, con cada sistema de corrección, TBC y sistema tradicional, con la nota del examen de preguntas abiertas. Por tanto tendrá mayor validez el sistema de puntuación que muestre una mayor correlación (coeficiente r de Pearson) con la nota obtenida en las preguntas abiertas.

4. Resultados

4.1. Elección adecuada del nivel de confianza por los estudiantes

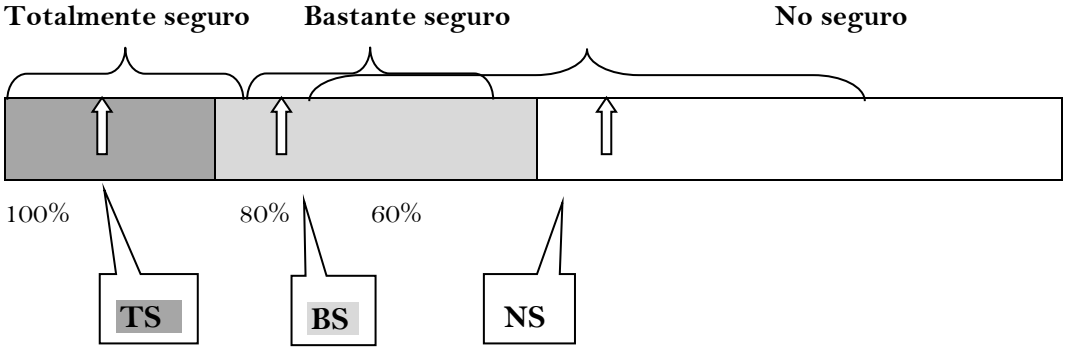
El porcentaje de respuestas correctas en cada nivel se recoge en el cuadro 5.

Cuadro 5. Porcentaje de aciertos con cada nivel de confianza

NIVEL DE SEGURIDAD	RESPUESTAS CORRECTAS
Aciertos *"totalmente seguro"*	93,48 %
Aciertos *"bastante seguro"*	73,14 %
Aciertos *"no seguro"*	52,03 %

Fuente: Elaboración propia.

Como se puede ver en la figura 2, los alumnos escogen su nivel de seguridad con criterios racionales adaptados a su nivel de conocimientos reales.



TS: *totalmente seguro*, BS: *bastante seguro*; NS: *no seguro*.

Figura 2. Intervalo de aciertos recomendado para cada nivel de confianza: Relación entre el porcentaje de aciertos y la opción de confianza escogida
Fuente: Elaboración propia.

4.2. Estudio de la consistencia interna del test con puntuación basada en el grado de seguridad

Para determinar la consistencia interna, se corrige cada examen por los dos procedimientos: las puntuaciones de cada pregunta corregida por TBC son las recogidas en el cuadro 3, y las del test tradicional en el cuadro 4. En el cuadro 6 se recogen los datos de α de Cronbach.

Cuadro 6. Fiabilidad de los exámenes según el sistema de corrección

AÑO Y CURSO	Nº DE SUJETOS	N.º DE ÍTEMS	α TBC	α TEST POR ACIERTOS	T
2009 Fisioterapia 1	25	44	0,56	0,30	1,18
2009 Fisioterapia 2	42	42	0,83	0,83	0,00
2009 Enfermería	40	32	0,80	0,78	0,31
2013 Enfermería	107	45	0,82	0,67	3,32**
2014 Fisioterapia	37	60	0,87	0,85	0,45
TOTAL	251		-	-	-

Nota: **p<0,05
Fuente: Elaboración propia.

En cuatro de los cinco casos el coeficiente alfa es más elevado y en otro son iguales.

Se calcula individualmente pues, de acuerdo con las recomendaciones de varios autores (Morales, 2008; Thomson, 1994), los coeficientes de fiabilidad de muestras distintas se deben presentar independientemente, no son sumables ni tiene sentido hacer la media de ellos, ya que la fiabilidad no es una propiedad intrínseca del instrumento (el examen, en este caso) sino de los resultados obtenidos al aplicar el instrumento a una muestra determinada (Morales, 2008). Por ello, se presentan los resultados en los seis grupos analizados independientemente.

4.3. Validez del TBC en comparación con el test por aciertos

Los resultados obtenidos correlacionando la nota obtenida con cada tipo de test con la obtenida en el examen de preguntas abiertas se recogen en el cuadro 7.

Cuadro 7. Correlación de la nota de los test con la obtenida en el examen de preguntas abiertas

TIPO DE EXAMEN	CORRELACIÓN CON LA NOTA EN *PREGUNTAS ABIERTAS*
Test basado en confianza (TBC)	0,705*
Test basado en número de aciertos	0,672

Notas: * $p<0.05$; 215 sujetos. La prueba de la t de Student se hizo comparando los valores de la r de Pearson obtenidos en los dos test, utilizando el programa SISA, online statistical analisis.
Fuente: Elaboración propia.

La correlación entre las puntuaciones del TBC con el examen de preguntas abiertas es mayor (0,70 vs0,67). Aunque la diferencia de correlaciones es pequeña, tiene significación estadística.

5. Discusión y conclusiones

La educación tradicional resalta el conocimiento, el conocimiento cierto y seguro, y obvia el reconocimiento de las incertidumbres y el manejo por el alumno de éstas. Pero en el mundo real, las incertidumbres, la debilidad de las evidencias, las dudas, son más frecuentes que las certezas. En nuestra sociedad, saturada de información, el reconocimiento de las incertidumbres y de la ausencia de evidencias es, con frecuencia, más importante que la posesión del conocimiento en sí (Gardner-Medwin, 2011). Y cuando reconocemos la incertidumbre, es necesario ponderarla, cuantificarla, para poder manejarla intelectualmente. La cuestión de la seguridad/incertidumbre en el propio conocimiento es de importancia trascendental en el proceso de aprendizaje; como el conocimiento suele tener un porcentaje de incertidumbre, no hay un conocimiento adecuado sin la destreza necesaria para reconocer y manejar las dudas.

Con el sistema de test basado en la confianza se espera que el alumno adquiera mejores estrategias de estudio y una mayor conciencia de sus lagunas formativas. Ante cada pregunta del test ya no importa sólo *acertar* o *fallar*, sino *saber*, *estar inseguro* o *no saber*. Con el TBC los alumnos aprenden a reflexionar sobre su saber, a preguntarse el porqué de los conceptos aprendidos: no basta un recuerdo inmediato desvinculado. A esta capacidad de juzgar el propio saber, Kleitman y Stankov (2001) la denominaron "automonitorización", y la definieron como la habilidad para aprehender y juzgar la calidad del propio trabajo cognitivo. La medida de la confianza del estudiante es una medida ajustada y precisa de la "automonitorización", por lo que incluirla en el proceso evaluativo parece pertinente, fundamental en el aprendizaje autorregulado (Sharma y Bewes, 2011; Panadero y Alonso-Tapia, 2014). Por este motivo se ha adoptado el sistema de test con

puntuación basada en la confianza. Pero ante un nuevo sistema de evaluación del que no existe experiencia en nuestro medio, cabe preguntarse si el alumno se adapta adecuadamente a esta innovación, si escoge racionalmente su opción de seguridad, si el sistema proporciona información válida del aprendizaje del estudiante y si al aplicar un sistema de examen más complejo y que no valora únicamente el mero conocimiento, los resultados del examen no pierden fiabilidad.

La presente investigación pone de manifiesto que los alumnos escogen su nivel de confianza con criterios racionales, en correspondencia con sus conocimientos reales. Cuando escogen el máximo nivel de confianza tienen respuestas correctas en el 93% de los casos y cuando optan por el nivel mínimo de confianza su porcentaje de respuestas correctas baja hasta el 52% de los casos. Así pues, los estudiantes aprenden a utilizar correctamente este sistema de examen, valorando de forma adecuada su nivel de confianza.

En el estudio de la validez, la correlación de las notas obtenidas mediante la puntuación de la confianza con la variable-criterio (examen de preguntas abiertas), superior en el test basado en la confianza con diferencia significativa a la obtenida con el sistema tradicional de corrección, permite afirmar que utilizar el sistema TBC mejora la validez de los resultados del examen frente al sistema de puntuación tradicional del test, que solo tiene en cuenta el número de aciertos.

La determinación de la consistencia interna muestra que el TBC es un sistema de examen que tiene mayor fiabilidad que la puntuación por aciertos en cuatro de los cinco casos estudiados, mientras que en el restante la validez es igual con ambos exámenes. Posiblemente el reducido número de alumnos en la mayoría de las pruebas (entre 25 y 42), dificulta que las diferencias puedan alcanzar significación estadística, mientras que, en el examen en el que participaron más estudiantes, 107, sí se encuentran diferencias estadísticamente significativas. Nosotros consideramos, al plantear el presente trabajo, que sería suficiente con demostrar que la consistencia interna no se reducía al utilizar el TBC, pero hemos comprobado que, de hecho aumenta en la mayoría de los casos. Aplicar la medida de la confianza a la simple medida del conocimiento mejora, en nuestra experiencia, la fiabilidad del examen.

Esto implica que el valorar dos conceptos aparentemente diferentes, como son *conocimientos* y *confianza*, no reduce ni la validez ni la consistencia interna del TBC en comparación con un examen –el test por aciertos- que solo mide conocimiento, sino que las aumenta. Esto establece un fuerte vínculo entre ambos conceptos, los conocimientos y la confianza; ambos son aspectos de un mismo factor, al que podríamos denominar "dominio de la materia". Sin conocimientos no hay, evidentemente, dominio de la materia, pero sin confianza tampoco. El test basado en la confianza junta ambos parámetros, conocimiento y confianza.

El TBC ha demostrado en nuestro medio docente ser un excelente procedimiento de evaluación, con evidencias de validez y fiabilidad, útil para el profesor y para el estudiante, que aporta ventajas frente al procedimiento habitual de puntuación de los exámenes de tipo test, basado únicamente en el número de aciertos. Consideramos que sería conveniente replicar el experimento con una muestra más amplia.

Referencias

Ahlgren, A. (1970). *A hand-scoring system for confidence-weighted scores.* Washington, DC: Department of Health, Education & Welfare, Office of Education.

Barr, D. A. y Burke, J. R. (2013). Using confidence-based marking in a laboratory setting: A tool for student self-assessment and learning. *The Journal of Chiropractic Education*, *27*(1), 21-26. https://doi.org/10.7899/JCE-12-018

Ben–Shakhar, G. y Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement*, *28*(1), 23-35. https://doi.org/10.1111/j.1745-3984.1991.tb00341.x

Biggs, J. B. (1993). What do inventories of students' learning process really measure? A theoretical review and clarification. *British Journal of Educational Psychology*, *63*(1), 3–19. https://doi.org/10.1111/j.2044-8279.1993.tb01038.x

Bloom, B. S. (1956). *Taxonomy of educational objectives: the classification of educational goal S: Handbook I, cognitive domain.* Londres: Longman Group.

Budescu, D. y Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement*, *30*(4), 277-91. https://doi.org/10.1111/j.1745-3984.1993.tb00427.x

Chang, C. Y. y Cheng, W. Y. (2008). Science achievement and students' self-confidence and interest in science: A Taiwanese representative sample study. *International Journal of Science Education*, *30*(9), 1183-200. https://doi.org/10.1080/09500690701435384

Ebel, R. L. (1968). Valid confidence testing: Demonstration kit. *Journal of Educational Measurement*, *5*(4), 353–354.

Entwistle, N., Hanley, M. y Ratcliffe, G. (1979). Approaches to learning and levels of understanding, *British Journal of Educational Research*, *5*(1), 99-114. https://doi.org/10.1080014119279005011O

Frary, R. B. (1982). A simulation study of reliability and validity of multiple–choice test scores under six response–scoring modes. *Journal of Educational Statistics*, *7*(4), 333-51. https://doi.org/ 10.3102/10769986007004333

Gardner, W. C. (septiembre, 1969). The use of confidence testing in the Academic Instructor Course. Comunicación presentada en el *Annual Conference of the Military Testing Association*, Nueva York.

Gardner-Medwin, A. R. (1995). Confidence assessment in the teaching of basic science. *Association for Learning Technology Journal*, *3*(1), 80-85. https://doi.org/10.3402/ rlt.v3i1.9597

Gardner-Medwin, A. R. (1998). Updating with confidence: Do your students know what they don't know?. *Health Informatics*, *4*, 45-46.

Gardner–Medwin, A. R. (2006). Confidence-based marking towards deeper learning and better exams. En C. Bryan y K. Clegg (Eds.), *Innovative Assessment in Higher Education* (pp. 141-149). Londres: Routledge.

Gardner–Medwin, A. R. (2011). Reasonable doubt: uncertainty in education, science and law. *Proceedings of the British Academy*, *171*, 465-83.

Gardner-Medwin, A. R. y Curtin, N. (2007). *Certainty-Based Marking (CBM) for reflective learning and proper knowledge assessment.* Recuperado de http://www.ucl.ac.uk/lapt/REAP_cbm.pdf

Gardner-Medwin, A. R. y Gahan, M. (julio, 2003). Formative and summative confidence-based assessment. Comunicación presentada en el *7th International Computer–Aided Assessment*

Conference, Loughborough, Reino Unido. Recuperado de: https://tmedwin.net/~ucgbarg/tea/caa03a.pdf

Hambleton, R. K., Roberts, D. M. y Traub, R. R. (1970). A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. J*ournal of Educational Measurement*, *7*(2), 75-90. https://doi.org/10.1111/j.1745-3984.1970.tb00698.x

Hassmen, P. y Hunt, D. P. (1994). Human self–assessment in multiple–choice testing. *Journal of Educational Measurement*, *31*(2), 149-60. https://doi.org/10.1111/j.1745-3984.1994.tb00440.x

Hevner, K. (1932). A method of correcting for guessing in true–false tests and empirical evidence in support of IT. *The Journal of Social Psychology*, *3*(3), 359-62. https://doi.org/10.1080/00224545.1932.9919159

Hunt, D. P. (2003). The concept of knowledge and how to measure it. *Journal of Intellectual Capital*, *4*(1), 100-13. https://doi.org/10.1108/14691930310455414

Kansup, W. y Hakstian, A. R. (1975). A comparison of several methods of assessing partial knowledge in multiple choice tests: I. Scoring procedures. *Journal of Educational Measurement*, *12*(4), 219-30. https://doi.org/10.1111/j.1745-3984.1975.tb01024.x

Khan, K. S., Davies, D. A. y Gupta, J. K. (2001) Formative self–assessment using multiple true–false questions on the internet: Feedback according to confidence about correct knowledge. *Medical Teacher*, *23*(2),158-63. https://doi.org/10.1080/01421590031107

Kleitman, S. y Stankov, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Journal of Applied Cognitive Psychology*, *15(3)*, 321-341. https://doi.org/10.1002/acp.705

Luetsch, K. y Burrows, J. (2016). Certainty rating in pre-and post-tests of study modules in an online clinical pharmacy course - A pilot study to evaluate teaching and learning. *BMC Medical Education*,*16*(1), 267-291. https://doi.org/10.1186/s12909-016-0783-1

Marton, F. y Säljo, R. (1976a). On qualitative differences of learning (I): outcome and process. *British Journal of Educational Psychology*,*46*(1), 4-11. https://doi.org/10.1111/j.2044-8279.1976.tb02980.x

Marton, F. y Säljo, R. (1976b). On qualitative differences in learning II: outcome as a function of the learner's conception of the task. *British Journal of Educational Psychology*, *46*(2), 115–27. https://doi.org/10.1111/j.2044-8279.1976.tb02304.x

Moore, D.A. y Healy, P.J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502-517.https://doi.org/10. 1037/0033-295X.115.2.502

Morales, P. (2006). *Las pruebas objetivas: normas, modalidades y cuestiones discutidas.* Recuperado dehttp://www.upcomillas. Es/personal/peter/otrosdocumentos /PruebasObjetivas. Pdf.

Morales, P. (2008). *Estadística aplicada a las Ciencias Sociales.* Madrid: Universidad Pontificia Comillas.

Morony, S., Kleitman, S., Lee, Y. P. y Stankov, L. (2013). Predicting achievement: confidence versus self-efficacy, anxiety, and self-concept in Confucian and European countries. *International Journal of Educational Research*, *58*, 79-96. https://doi.org/10.1016/j.ijer.2012.11.002

Panadero, E. y Alonso-Tapia, J. (2014). How do students self-regulate? Review of Zimmerman's cyclical model of self-regulated learning. *Anales de Psicología*, *30*(2), 450-462. https://doi.org/10.6018/analesps.30.2.167221

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching and assessing. *Theory into Practice*, *41*(4), 219-25. https://doi.org/10.1207/s15430421tip4104_3

Rippey, R. M. (1978), Interactive confidence test scoring and interpretation. *Educational and Psychological Measurement*, *38*(1), 153-57. https://doi.org/10.1177/001316447803800122

Sharma, M. D. y Bewes, J. (2011). Self-monitoring: Confidence, academic achievement and gender differences in Fhysics. *Journal of Learning Design*, *4*(3), 1-13. https://doi.org/10.5204/jld.v4i3.76

Stanger-Hall, K.F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. *CBE Life Scientific Education.11*(3), 294-306. https://doi.org/10.1187/cbe.11-11-0100

Stankov, L. (2000). Complexity, metacognition and fluid intelligence. *Intelligence*, *28*(2), 121-43 . https://doi.org/10.1016/S0160-2896(99)00033-1

Stankov, L., Lee, J. y Paek, J. (2009). Realism of confidence judgments. *European Journal of Psychological Assessment*, *25*, 123-30. https://doi.org/10.1027/1015-5759.25.2.123

Stankov, L., Lee, J., Luo, W. y Hogan, D. J. (2012). Confidence: A better predictor of academic achievement self-efficacy, self-concept and anxiety?. *Learning and Individual Differences*, *22*(6), 747-58. https://doi.org/10.1016/j.lindif.2012.05.013

Thomson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*,837-47.

Urosa, B. (1995). *La adivinación en las pruebas objetivas: Alternativas a la fórmula clásica de corrección.* Tesis Doctoral, Universidad Pontificia Comillas.

Breve Cv de los autores

Carlos Valencia Rodríguez

Doctor en Ciencias de la Educación, licenciado en Medicina. Profesor Agregado de la Universidad Pontificia Comillas, imparte clases en Ciencias Básicas de la Salud en la Escuela de Enfermería y Fisioterapia "San Juan de Dios". Está especializado en investigación sobre técnicas de aprendizaje en Ciencias de la Salud. ORCID ID: https://orcid.org/0000-0003-3146-6615. E-mail: cvalencia@comillas.edu

Yolanda Ortega Latorre

Doctora en Ciencias Humanas y Sociales, licenciada en Farmacia. Profesora Colaboradora de la Universidad Pontificia Comillas, imparte clases Ciencias Básicas de la Salud en la Escuela de Enfermería y Fisioterapia "San Juan de Dios". Participa en proyectos de investigación sobre automedicación en adolescentes y estilos de vida saludables y en proyectos de investigación sobre técnicas de evaluación en el aprendizaje en Ciencias de la Salud. ORCID ID: https://orcid.org/0000-0002-0107-295X. E-mail: yol.ortega@comillas.edu

Paloma Huerta Cebrián

Doctora en Farmacia. Profesora Agregada de la Universidad Pontificia Comillas, imparte clases Ciencias Básicas de la Salud en la Escuela de Enfermería y Fisioterapia "San Juan de Dios". Participa en proyectos de investigación sobre automedicación en adolescentes y estilos de vida saludables, así como en proyectos de investigación sobre técnicas de evaluación en el aprendizaje en Ciencias de la Salud. ORCID ID: https://orcid.org/0000-0001-9223-4549. E-mail: phuerta@comillas.edu