Inv Ed Med. 2016;5(17):55–63

Investigación en Educación Médica

http://riem.facmed.unam.mx

ARTÍCULO DE REVISIÓN

Reflexiones sobre adivinar en preguntas de opción múltiple y cómo afecta el resultado del examen

Alma Jurado-Núñez[a] **e Iwin Leenen**[b,*]

[a] *Programa de Apoyo y Fomento a la Investigación Estudiantil (AFINES), Facultad de Medicina, Universidad Nacional Autónoma de México, México D.F., México*
[b] *Departamento de Evaluación, Secretaría de Educación Médica, Facultad de Medicina, Universidad Nacional Autónoma de México, México D.F., México*

Recibido el 18 de enero de 2015; aceptado el 27 de julio de 2015
Disponible en Internet el 9 de septiembre de 2015

PALABRAS CLAVE
Exámenes de opción múltiple;
Adivinar;
Teoría Clásica de los Tests;
Teoría de Respuesta al Ítem;
México

Resumen Los exámenes de opción múltiple (EOM) son la herramienta más difundida en educación médica, pero su utilidad está supeditada a la confiabilidad del instrumento y la validez de las inferencias que emanan de la medición. La posibilidad de adivinar, inherente al formato de evaluación, puede introducir varianza irrelevante a la medición y reducir la representación del rasgo latente en la calificación del examen por diferencias individuales respecto a *educated guessing*, *testwiseness* y la tendencia a adivinar. En este artículo se presentan brevemente las características generales de la Teoría Clásica de los Tests (TCT) y la Teoría de Respuesta al Ítem (TRI) y su abordaje al problema de adivinar. Asimismo, se propone un modelo teórico dentro de la TCT que integra los mecanismos que afectan la adivinación y se determina la variación de la probabilidad de aprobar un EOM, en función de ciertos supuestos respecto a adivinar a través de un análisis teórico dentro de un modelo TRI. Es posible concluir que algunas características de los ítems propician la adivinación, y cuando ésta ocurre se encuentran inmersas diversas variables, relacionadas o independientes, del rasgo que se pretende medir, que determinan la magnitud de su efecto.
Derechos Reservados © 2015 Universidad Nacional Autónoma de México, Facultad de Medicina. Este es un artículo de acceso abierto distribuido bajo los términos de la Licencia Creative Commons CC BY-NC-ND 4.0.

KEYWORDS
Multiple-choice tests;
Guessing;
Classical Test Theory;

Insights into guessing in multiple choice questions and its effect in the assessment outcome

Abstract Multiple-choice tests (MCT) are the most employed assessment tool in medical education; however, its use is limited to the instrument reliability and validity of the inferences made

* Autor para correspondencia: Departamento de Evaluación, Secretaría de Educación Médica. Av. Universidad N° 3000, Edif. B, 3er piso, C.U., C.P. 04510, México D.F., México. Tel.: +56-23-23-00 Ext. 43034.
Correo electrónico: iwin.leenen@gmail.com (I. Leenen).
La revisión por pares es responsabilidad de la Universidad Nacional Autónoma de México.

http://dx.doi.org/10.1016/j.riem.2015.07.004
2007-5057/Derechos Reservados © 2015 Universidad Nacional Autónoma de México, Facultad de Medicina. Este es un artículo de acceso abierto distribuido bajo los términos de la Licencia Creative Commons CC BY-NC-ND 4.0.

Item Response Theory;
Mexico

upon the measurement. Guessing, an inherent element of this evaluating tool, may introduce construct-irrelevant variance and reduce the load of the latent trait in the score of the exam due to individual differences regarding educated guessing, testwiseness and guessing tendency. This article presents an overview of Classical Test Theory (CTT) and Item Response Theory (IRT), including a discussion of how each theory addresses the guessing phenomenon. With respect to the latter, we propose a theoretical model that integrates factors related to guessing within the CTT framework. We further include a theoretical analysis, which displays the variation of the probability of passing a MCT reliant on certain assumptions regarding guessing that are akin to a particular IRT model. In conclusion, various features of the items increase the likelihood of guessing, and, when guessing takes place, the magnitude of its effect is determined by some variables that can be dependent or independent of the latent trait.
All Rights Reserved © 2015 Universidad Nacional Autónoma de México, Facultad de Medicina. This is an open access item distributed under the Creative Commons CC License BY-NC-ND 4.0.

Introducción

Los exámenes de opción múltiple (EOM) son la herramienta más utilizada en educación médica al hacer posible una evaluación objetiva, estandarizada, costo-efectiva y eficiente[1–3]. Un ítem de opción múltiple (IOM) desarrollado de forma correcta puede evaluar desde memoria hasta procesos cognitivos superiores como razonamiento clínico y toma de decisiones. Otras ventajas incluyen: *a)* un muestreo más extenso del contenido por evaluar, y *b)* una administración y calificación en poco tiempo, aun en grandes poblaciones de estudiantes[4]. Esta forma de evaluación se ha utilizado a gran escala internacionalmente y ha sido un tema central en investigación en educación desde el siglo pasado. Lo que se espera de cualquier método de evaluación educativa es que proporcione una medición lo más certera posible del rasgo latente que se pretende medir. La validez del método es especialmente trascendente cuando se trata de exámenes sumativos, donde el puntaje se utiliza para tomar decisiones sobre los sustentantes (en educación médica, concretamente, de éste puede depender la emisión de un título o el acceso a un curso de especialidad). Por ello, es indispensable que los EOM, como instrumento de medición, sean lo más confiables, válidos y justos posible.

Respecto a estas tres características, surge un elemento inherente a este formato de evaluación que siempre se ha cuestionado: la posibilidad de adivinar. Existe la posibilidad de que los sustentantes sumen puntos que no representen dominio sobre el tema e incluso, al menos teóricamente, que aprueben un examen a expensas de una ventaja del propio formato. Esta posibilidad se presenta aunque se adivine a ciegas. Sin embargo, pocas veces el sustentante tiene que adivinar a ciegas. Por un lado, los EOM suelen incluir ítems con deficiencias en su construcción, que permiten descartar una o más opciones o inferir la respuesta correcta, incluso con un bajo nivel de conocimiento[5–7]. Este fenómeno se conoce como *testwiseness*[8]. Por otro lado, es común que, con base en un conocimiento parcial, el sustentante pueda descartar uno o más distractores de un IOM, lo cual se ha denominado *informed guessing* o *educated guessing*[9–11]. En general, adivinar constituye una fuente de error de medición que puede convertirse en sesgo estadístico[12]. Por lo tanto, la decisión sobre el punto de corte entre aprobar y reprobar un EOM debería tomar en cuenta que un porcentaje de los aciertos no refleja el dominio del tema o contenido evaluado por la pregunta, sino que se debe al azar.

Para el análisis de pruebas en general (y EOM en específico), existen dos enfoques psicométricos principales: la Teoría Clásica de los Tests (TCT) y la Teoría de Respuesta al Ítem (TRI). Ambos modelos tienen como objetivo realizar inferencias sobre uno o más rasgos latentes, imposibles de observarse directamente, a partir de las respuestas a las preguntas. Por otro lado, sus supuestos subyacentes difieren de forma notable, incluyendo la forma en que consideran y enfrentan el problema de acertar preguntas por adivinación.

El presente artículo tiene dos objetivos: primero, revisamos brevemente las características de la TCT y la TRI, en general (para una revisión más extensa, véase Leenen[13]) y cómo cada teoría ha abordado el problema de adivinar. Al respecto, incluimos *a)* un modelo que integra los mecanismos que afectan la adivinación con los conceptos claves de la TCT, y *b)* una breve descripción de algunos modelos TRI específicos para EOM. Segundo, investigamos, a partir de un análisis teórico enmarcado dentro del modelo Nedelsky, cómo la probabilidad de aprobar un EOM varía en función de ciertos supuestos o estrategias respecto a adivinar. Los resultados de esta investigación aportan evidencia sobre la (baja) plausibilidad de acreditar un examen a expensas de adivinar a ciegas o con base en *testwiseness* o *educated guessing* exclusivamente.

Dos marcos psicométricos y cómo tratan al problema de la adivinación

Conceptos generales de la Teoría Clásica de los Tests

Los supuestos y procedimientos de la TCT son relativamente simples y consideran la prueba en su totalidad. Distingue entre dos factores que componen la puntuación observada (comúnmente representada por X) en el examen: la puntuación verdadera (V) y el error de medición (E):

$$X = V + E \tag{1}$$

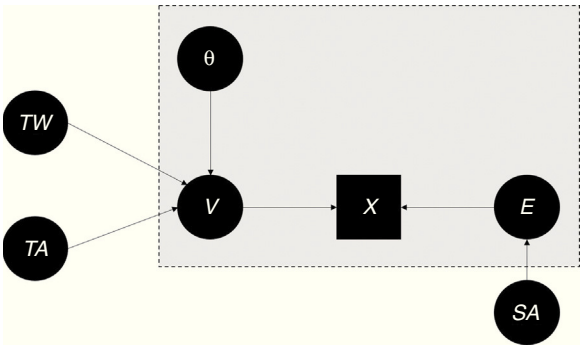

Figura 1 Modelo teórico que integra suerte al adivinar (SA), *testwiseness* (TW), tendencia a adivinar (TA) con los conceptos centrales de la TCT, puntuación verdadera (*V*), rasgo latente (θ), error de medición (*E*) y la puntuación obModelo teórico que integra suerte al adivinar (SA), *testwiseness* (TW), tendencia a adivinar (TA) con los conceptos centrales de la TCT, puntuación verdadera (*V*), rasgo latente (θ), error de medición (*E*) y la puntuación observada (*X*).

La TCT parte de la concepción teórica e hipotética de que el examen se puede replicar un gran número de veces y que, en cada réplica, las características esenciales (por ejemplo, ciertos contenidos que se deben cubrir según la tabla de especificaciones y el nivel taxonómico de las preguntas) permanecen, mientras que las características incidentales (como las preguntas concretas y las circunstancias de aplicación) varían. La puntuación verdadera, por definición, recoge todos los efectos sistemáticos, es decir, de los factores que permanecen entre réplicas. Ésta incluye el efecto del constructo latente que se pretende medir (habitualmente representado por θ) y de otros factores, como las estrategias para adivinar en caso de incertidumbre o la familiaridad con el formato de respuesta. Como se explicará posteriormente, θ es central al considerar la validez del examen, mientras que los demás factores generan varianza irrelevante para el constructo[14,15]. El error de medición reúne los efectos no sistemáticos, o sea aquellos que varían entre las réplicas hipotéticas, como por ejemplo, la mala suerte de equivocarse por la presencia de una distracción (como por ruido excesivo). La parte gris de la figura 1 representa las cuatro variables principales en el modelo clásico y su relación.

Sigue de la ecuación (1) que las diferencias entre sustentantes respecto de su puntuación observada reflejan tanto diferencias verdaderas como diferencias incidentales; la confiabilidad en la TCT se define como la proporción de diferencias verdaderas en las diferencias observadas. Se puede entender la confiabilidad en términos del efecto relativo que tienen los factores *V* y *E* sobre (la varianza en) *X*. El concepto estrechamente relacionado del error estándar de medición se suele usar para derivar un intervalo de confianza para la puntuación verdadera con base en la puntuación observada[16,17].

El concepto de validez[15,18,19] se refiere al efecto que tiene el constructo latente θ sobre *X*. Como se muestra en la figura 1, éste se manifiesta a través de *V* y deja claro que la validez depende de la confiabilidad de un instrumento[16].

Adivinar en la Teoría Clásica de los Tests

Suerte al adivinar, *testwiseness*, *educated guessing* y la tendencia a adivinar

Aunque la TCT y sus conceptos centrales refieren a la prueba en su totalidad, es conveniente analizar el proceso interno que ocurre en el sustentante al enfrentarse con un IOM. Al respecto, es razonable suponer que éste realiza un análisis tácito de las *m* opciones de respuesta, el cual resulta en una valoración sobre la plausibilidad de cada opción. Con base en estas valoraciones, elegirá entre las alternativas ofrecidas, o bien, dejará la pregunta sin responder. La adivinación ocurre si el sustentante decide contestar la pregunta, a pesar de que su análisis no haya conducido a la certidumbre respecto a una de las opciones.

Formalmente, estas valoraciones desembocan en *m* + 1 probabilidades: una probabilidad para cada opción de que el sustentante la elija, y otra de dejar la pregunta sin responder. En el caso de que la probabilidad asociada con la opción correcta sea diferente de 0 o 1, la puntuación en esta pregunta (y, por lo tanto, la puntuación observada *X* en el test) está influida por un factor aleatorio que denominamos *suerte al adivinar* (SA). La SA tiene un efecto no sistemático, por lo cual es parte del error de medición *E* (véase la figura 1). Obviamente, la SA interviene cuando el sustentante adivina a ciegas. En este caso, por definición, las probabilidades asociadas con las *m* opciones de respuesta son idénticas. Sin embargo, es vital enfatizar que la SA también interviene, por ejemplo, cuando el sustentante sabe eliminar algún distractor y adivina entre las opciones restantes.

En la introducción se mencionaron dos conceptos relacionados con la adivinación: *testwiseness* (TW) y *educated guessing* (EG). Ambos se relacionan con el proceso de valorar la plausibilidad de las opciones de los IOM al implicar valoraciones distintas que, por ende, llevan a probabilidades no uniformes de que se elija cada opción. En el caso de TW, las valoraciones de las *m* opciones de respuesta se deben a características relacionadas con aspectos gramaticales y sintácticos del ítem, no con el rasgo θ. EG, por otro lado, entra en escena cuando la diferenciación entre opciones de respuesta se debe al rasgo latente. Es importante señalar que los efectos de TW y EG no necesariamente incrementan la probabilidad de acertar el ítem (en comparación con la probabilidad de acertarla al adivinar a ciegas). Puede ser, por ejemplo, que en un ítem concreto, tengan un efecto engañoso, esto es, que lleven al sustentante a bajar la probabilidad que le asigna a la respuesta correcta.

Aunque estos dos conceptos a menudo se consideran de forma simultánea[9,20,21], es básico reconocer que tienen una concepción psicológica distinta. TW refiere a una capacidad del sustentante, es decir, es un *constructo* que ejerce un efecto sistemático sobre *X* (véase la flecha directa de TW hacia *V* en la figura 1), pero sin relación directa con θ (nótese la ausencia de una flecha entre TW y θ). A diferencia de TW, EG es una *conducta* que se manifiesta cuando θ influye en el proceso de la valoración de las opciones de respuesta, sin que éste lleve al sustentante a la identificación certera de una opción como la correcta. Debe ser claro que EG depende totalmente de θ y que está típicamente asociado con niveles intermedios de θ, donde el conocimiento parcial del sustentante es suficiente para, por ejemplo, descartar

un distractor, pero insuficiente para reconocer la respuesta correcta. Puesto que EG es una conducta subordinada totalmente a θ, está implícitamente presente en la flecha que indica el efecto de θ sobre *V* en la figura 1.

La figura 1 incluye un tercer factor que afecta *V*: la *tendencia a adivinar* (TA). Contrario a TW y EG, la acción de la TA no conlleva una distinción entre las *m* opciones de respuesta, sino afecta la probabilidad de dejar o no en blanco el ítem en caso de incertidumbre. Especialmente cuando se aplica una fórmula de corrección (véase abajo), los sustentantes difieren respecto a la decisión de dejar IOM sin contestar[20]. La aversión al riesgo y la concepción respecto a la práctica de adivinar son algunos factores que inciden sobre la magnitud de la TA. Por ejemplo, alguien puede opinar que es incorrecto sumar puntos a través de la suerte o que adivinar distorsiona la calificación en el examen; por tanto, la magnitud del efecto de la TA sería baja. Cabe mencionar que el efecto de la TA, al referirse a una tendencia general del sustentante hacia la incertidumbre en los IOM, es sistemático sobre *V*; se considera una variable moderadora que influye en cómo θ y TW afectan a *V*.

La regla para calificar

Típicamente, la calificación en un EOM se calcula como el número de ítems en los que el sustentante marcó la respuesta correcta. Esta regla se conoce como *number right scoring* (NRS) y, bajo la premisa que el sustentante desea maximizar su puntuación, siempre es una invitación a adivinar.

Con el fin de obtener una calificación menos afectada por la adivinación, en algunos contextos se aplica una *fórmula de corrección* al número de respuestas correctas. Esta regla alternativa castiga—es decir, resta puntos por—las respuestas incorrectas. Si se decide aplicar la fórmula, es esencial mencionarle a los sustentantes, previo al inicio de la prueba, que existirá dicha penalización y que se aplicará a los ítems con respuesta incorrecta (dejando sin penalización los ítems sin responder). La penalización suele depender del número *m* de opciones de respuesta y en la mayoría de las aplicaciones consiste en restar un valor de $1/(m-1)$ por cada respuesta incorrecta. Con esta corrección, la puntuación esperada (en el examen o en cada IOM) bajo el supuesto de adivinar a ciegas es cero[22]. No obstante, el efecto más trascendente de la fórmula se debe al discurso precautorio y a su potencial de lograr que los sustentantes se abstengan de intentar adivinar, más que a la deducción de puntos[12,22]. Así, la fórmula puede reducir la varianza del error en comparación con la calificación obtenida por el NRS, ya que este último cuasi fuerza a los sustentantes a adivinar cuando desconocen la respuesta.

Sin embargo, la fórmula de corrección ha sido criticada porque esta regla introduce varianza sistemática irrelevante para el constructo, debido a que los sustentantes reaccionan de forma diferente a la posible penalización[22–24]. Al prevenir a los estudiantes sobre la fórmula, se introducen nuevos elementos, como su personalidad y principalmente su actitud hacia el riesgo, que afectan la estrategia de resolución del examen y distorsionan el puntaje final. En otras palabras, la propensión a tomar riesgos y la tendencia a adivinar se vuelven más trascendentes para los parámetros del examen que la propia fórmula de corrección[25].

Es interesante plantear el debate entre defensores y oponentes de la fórmula de corrección dentro del modelo que se presentó en la figura 1. El argumento a favor enfatiza la disminución del efecto de la SA sobre la puntuación observada, la cual beneficia la confiabilidad del examen. El argumento en contra resalta que la fórmula fortalece la influencia de la TA sobre *V* a expensas de θ y, por ende, atenúa la validez del examen. Según los detractores de la fórmula, la intención de que se reponga la confiabilidad perdida en adivinar no se logra, y se compromete más la validez del instrumento por la varianza irrelevante agregada a la medición.

La Teoría de Respuesta al Ítem

La TRI constituye una familia de modelos que formalizan el proceso de responder a un ítem. En cualquier modelo TRI es central la *función característica*, la cual relaciona las características de los ítems con los rasgos latentes de los examinados a fin de precisar las probabilidades de observar ciertas (categorías de) respuestas. La TRI tiene un fundamento matemático más robusto, con supuestos más fuertes y precisos e interpretaciones más claras en comparación con la TCT[26,27]. Para conocer más respecto a las diferencias entre la TCT y la TRI, véase Leenen[13], Hambleton y Jones[28] y Erguven[29].

Dentro de la TRI, se encuentra una diversidad de modelos que difieren en el número de parámetros (cuantificaciones de las características de personas e ítems), y en la función característica que los une para llegar a la afirmación probabilística de acertar el ítem o marcar cierta opción. Por ejemplo, el modelo de Rasch[30] —uno de los pioneros de la TRI— supone *a)* un parámetro para cada ítem (β, su dificultad); *b)* un parámetro por persona (su nivel θ de habilidad), y *c)* que la probabilidad de que una persona acierte un ítem (el modelo solo considera dos categorías de respuesta, correcta o incorrecta) crece monótonamente conforme la diferencia $\theta - \beta$ aumenta. El modelo es unidimensional y, como muchos otros modelos de la TRI, asume independencia local.

Adivinar en la Teoría de Respuesta al Ítem

Una propiedad importante del modelo de Rasch es que la probabilidad de acertar un ítem se acerca a cero conforme el nivel θ de la persona disminuye. Por ello, el modelo no se considera muy apropiado para el análisis de EOM, puesto que incluso una persona totalmente ignorante tiene una probabilidad sustancialmente superior a cero de acertar el ítem. Para responder a esta inconveniencia, se han propuesto modelos TRI alternativos que explícitamente consideran la adivinación en los IOM. A continuación, se describen brevemente algunos de éstos.

El modelo logístico de tres parámetros

El modelo TRI más popular y más común para el análisis de EOM es, indudablemente, el modelo logístico de tres parámetros (3PL)[31]. Este modelo incorpora explícitamente la posibilidad de adivinar, al incluir un parámetro (γ, de pseudoadivinación) para cada ítem que representa la probabilidad de acertarlo para personas de un nivel muy/infinitamente bajo. (Cabe mencionar que, a diferencia con el modelo de Rasch, el 3PL incluye otro parámetro

para cada ítem, α, su discriminación; sin embargo, éste no es relevante para este artículo.)

La interpretación más común del modelo 3PL supone un proceso en dos pasos: *a)* el sustentante analiza el ítem y, con base en el resultado de este análisis, *b)* provee la respuesta correcta (si la conoce) o adivina (si la desconoce). La probabilidad total de acertar el ítem, como está formalizado en la función característica del 3PL, es la suma de dos probabilidades: la de conocer la respuesta correcta, en cuyo caso acierta con certeza, y la probabilidad conjunta de desconocer la respuesta correcta y acertar por adivinación[32].

Un modelo Teoría de Respuesta al Ítem en el cual la adivinación depende de la persona

El 3PL restringe el parámetro de pseudoadivinación a ser dependiente del ítem y no de la persona que responde. Sin embargo, es poco plausible suponer que, si el estudiante desconoce la respuesta y adivina, la probabilidad de acertar el ítem es fija, o sea, que no depende de él. Más probable es, entonces, que utilice información parcial durante el proceso de adivinar. Para remediar este inconveniente, San Martin et al.[33] propusieron un modelo en el cual la probabilidad de acertar el ítem por adivinación depende en cierto grado de la variable latente θ.

El modelo de respuesta nominal y sus generalizaciones

Tanto el 3PL como el modelo de San Martín et al. unen los distractores y la no respuesta en una categoría, la ''respuesta incorrecta''. Esta práctica, aunque es muy común y refleja la costumbre de otorgar un punto a la respuesta correcta y cero puntos a cualquier respuesta incorrecta, implica pérdida de información. Posiblemente, tomar en cuenta *cuál* opción incorrecta eligió una persona puede llevar a una estimación más precisa de la θ de ésta, sobre todo para niveles relativamente bajos de la variable latente[34,35].

Se han propuesto varios modelos TRI que permiten analizar las múltiples categorías de respuesta en los IOMs; el primero y más conocido es el modelo de respuesta nominal[36], el cual parte del supuesto de que las m opciones de respuesta en un IOM ejercen diferentes grados de atracción sobre el sustentante. Si $a_1, a_2, \ldots, a_m$ son números positivos que cuantifican dicha atracción, entonces el modelo especifica que la probabilidad de escoger la opción j (donde j es un índice entre 1 y m) se da por:

$$\Pr(\text{opción}j) = \frac{a_j}{a_1 + a_2 + \cdots + a_m}. \qquad (2)$$

La fuerza de atracción es conceptualmente similar a la valoración de las opciones de respuesta en el modelo que se describió en la sección *Adivinar en la TCT*. Los valores $a_1, a_2, \ldots, a_m$ de la ecuación (2) son una función (es decir, dependen) tanto de los parámetros de las opciones de respuesta como de la persona.

El modelo de respuesta nominal comprende algunos detalles teóricamente menos deseables. Por ejemplo, la función característica implica que personas totalmente ignorantes siempre estarán atraídas a una opción específica, lo cual es poco plausible; es más verosímil que estas personas sean indiferentes entre las opciones ofrecidas. Los modelos de Samejima[37], Thissen y Steinberg[34], y más recientemente Suh y Bolt[21], remedian estos inconvenientes.

El modelo Nedelsky

El modelo Nedelsky[38] asimismo distingue entre los distintos distractores de un IOM, pero adopta unos supuestos más simples —y sobre todo psicológicamente diferentes— que los mencionados en el párrafo anterior. Supone un proceso que consiste en dos pasos. Primero, el sustentante realiza de forma independiente una evaluación de cada opción de respuesta; en el caso de que sea un distractor, la evaluación *posiblemente* le lleva a identificar la opción como incorrecta. La probabilidad de que esto ocurra depende de características de la opción (como su dificultad) y del nivel del sustentante en la variable latente (las personas difieren respecto a su habilidad para identificar distractores). Nótese que el modelo excluye la posibilidad de rechazar la opción correcta en este paso (se supone que en un IOM correctamente desarrollado, ningún nivel de θ puede llevar al rechazo de la opción correcta). En el segundo paso, el sustentante adivina a ciegas entre las opciones que no rechazó en el paso anterior. Si, por ejemplo, en un ítem de cuatro opciones de respuesta, el sustentante sabe identificar dos distractores en el primer paso, entonces en el segundo paso elegirá una de las dos opciones restantes con probabilidad de 0.50.

Nota final sobre los modelos Teoría de Respuesta al Ítem para ítem de opción múltiple

Para concluir esta sección, es importante resaltar que todos los modelos descritos, excepto el 3PL, al permitir que la variable latente θ intervenga en el proceso de respuesta cuando el sustentante adivina, incorporan y formalizan la posibilidad de EG. Al mismo tiempo, siendo modelos unidimensionales, excluyen explícitamente la posibilidad de que otros constructos, como TW o la TA, determinen la probabilidad de acertar el ítem o escoger cierta opción.

La probabilidad de aprobar un exámenes de opción múltiple

En esta sección, se presentan los resultados de un análisis teórico de la probabilidad de aprobar un EOM bajo seis escenarios diferentes. Como estándar de pase tomamos el criterio de obtener el 60% de la calificación máxima, lo cual es habitual en el sistema educativo de México. Todos los análisis refieren a exámenes de hasta 60 ítems, cada uno con cuatro opciones, y se pueden enmarcar dentro del modelo Nedelsky (véase la sección anterior). En particular, se supone que en cada ítem ocurre una de cuatro alternativas: el sustentante puede descartar 0, 1, 2, o los 3 distractores y, enseguida, adivina ciegamente entre las opciones restantes (sin dejar ítems sin contestar). Los escenarios difieren respecto de las probabilidades que se asocian a estas alternativas.

La tabla 1 presenta la distribución de probabilidades para cada escenario. El primer escenario corresponde con adivinar a ciegas todo el examen (en el 100% de los ítems no sabe descartar ningún distractor). En el segundo (y tercer) escenario, se supone que el sustentante puede descartar uno (o dos) distractores de cada ítem. Nótese que esto equivale a adivinar a ciegas en un examen donde los IOM tienen tres (o dos) opciones de respuesta. En contraste con los primeros tres escenarios, los últimos permiten que las cuatro

Tabla 1 Probabilidades de eliminar cierto número de distractores en una pregunta de cuatro opciones de respuesta para seis escenarios hipotéticos analizados bajo el modelo Nedelsky

Escenarios		Probabilidad de eliminar *x* distractores			
		x = 0	x = 1	x = 2	x = 3
1	Adivinar a ciegas	100%	0%	0%	0%
2	Descartar un distractor	0%	100%	0%	0%
3	Descartar dos distractores	0%	0%	100%	0%
4	Estudiante nivel bajo	10%	40%	20%	30%
5	Estudiante nivel medio	5%	30%	20%	45%
6	Estudiante nivel alto	1%	10%	19%	70%

alternativas ocurran en el mismo examen. La distribución de probabilidades en éstos responde a la experiencia de los autores al sustentar exámenes. Por ejemplo, el cuarto escenario especifica que el sustentante adivina a ciegas en el 10% de los ítems, descarta uno o dos distractores en el 40% y 20%, respectivamente, y que descarta los tres distractores en el 30% de los casos. Este escenario se puede relacionar con el desempeño de un estudiante de bajo nivel académico. Por otro lado, los dos últimos escenarios refieren más a estudiantes de nivel académico medio y alto, respectivamente.

En la figura 2, se muestra la probabilidad de aprobar en función del número de ítems que contiene el examen, para cada escenario y bajo dos reglas para calificar: *NRS* y aplicando la fórmula de corrección (restando $1/(m-1)$ por respuesta errónea). El gráfico superior izquierdo evidencia que la probabilidad de aprobar un examen con 10 ítems, adivinando a ciegas, es aproximadamente 2%, si la calificación se obtiene por NRS. Desde 20 ítems, la probabilidad de aprobar se vuelve despreciable (< 0.1%) y con 50 ítems aprobar un EOM a ciegas es prácticamente imposible. Con la fórmula de corrección, obviamente, se llega aún más rápido a la asíntota de 0.

Cuando el sustentante puede descartar un distractor en cada ítem, el patrón de probabilidades es similar: en particular, se evidencia que, para los exámenes comunes (es decir, de 50 ítems o más), la probabilidad de aprobar es despreciable. En cambio, los sustentantes que pueden identificar *dos* opciones incorrectas en cada ítem y adivinan a ciegas entre las dos restantes, tienen, en un examen de 50 ítems, una probabilidad de 10% de aprobar. Es importante resaltar dos implicaciones de este resultado: *a)* existe una probabilidad significativa (aunque baja) de que un estudiante apruebe un EOM sin poder identificar la respuesta correcta en ningún ítem, aún si contiene un número habitual de éstos, y *b)* exámenes donde los ítems tienen sólo dos opciones (como verdadero-falso) son susceptibles a ser aprobados adivinando a ciegas. Las implicaciones anteriores se refieren al caso de NRS; aplicando la corrección por adivinar las probabilidades se reducen a valores despreciables. Así, se ejemplifica la utilidad de la fórmula de corrección para evitar que aquellos que desconocen la respuesta correcta acumulen puntos.

Las gráficas correspondientes al Escenario 4, que corresponde a un estudiante de *bajo* nivel que no debería pasar el examen, muestran que la probabilidad de aprobar, sin la fórmula de corrección, no es inferior a 30%, incluso con 60 ítems. Si se aplica la penalización, la probabilidad se reduce a aproximadamente 2%. En las gráficas correspondientes al estudiante de nivel *medio*, se observa un efecto muy pronunciado de la fórmula de corrección: la evolución de la probabilidad de aprobar en función del número de ítems está totalmente supeditada a ésta. Con el NRS, la probabilidad de aprobar tiende a 1 conforme el número de ítems aumenta, mientras que con la fórmula tiende a 0. En un examen de 60 ítems, las probabilidades son de 87% (sin corrección) versus 32% (con corrección). El ejemplo evidencia que la decisión sobre la regla para calificar es transcendental. Bajo las condiciones del último escenario, correspondiente al sustentante de *alto* nivel, las probabilidades de aprobar el examen exceden 95% aún con un examen de solo 8 ítems, mientras no se aplique la fórmula de corrección. Este escenario muestra que la fórmula generalmente no perjudica al estudiante de alto nivel, siempre que contenga un número suficiente de ítems (a partir de 17 ítems, la probabilidad de aprobar el examen excede 90%).

Discusión y conclusiones

Tras el análisis teórico en la sección anterior, es preciso vincular los resultados con los abordajes de la adivinación según los dos marcos psicométricos. El Escenario 1, donde sólo sucede la adivinación a ciegas, ejemplifica la gran carga que ejercería el efecto de *SA* sobre la puntuación observada, ya que el resultado dependería directamente de éste. Sin embargo, la adivinación a ciegas es una práctica aislada, que difícilmente se sistematiza como estrategia de resolución, especialmente en exámenes de medianas o altas consecuencias. En los demás escenarios, al descartar una o más opciones, se introducen los efectos de TW y EG. Bajo la condición de unidimensionalidad de la TRI, la identificación de un distractor sólo podría deberse a EG; si el efecto de TW tiene una influencia significativa, las pruebas de ajuste típicamente llevan a un rechazo del modelo y se requieren variantes multidimensionales. Desde la TCT, mientras más influencia ejerce la acción de la TA y TW sobre *V*, mayor es la varianza irrelevante al constructo, y menos válidas resultan las inferencias realizadas sobre la puntuación observada.

Las propias características de los ítems son las que desencadenan más o menos efecto de EG y TW. Varios autores han identificado que la mayoría de los EOM en ciencias de la salud –especialmente en educación médica de pregrado y posgrado– albergan ítems que violan las directrices de construcción[7,10,39,40,45]. En general, dos

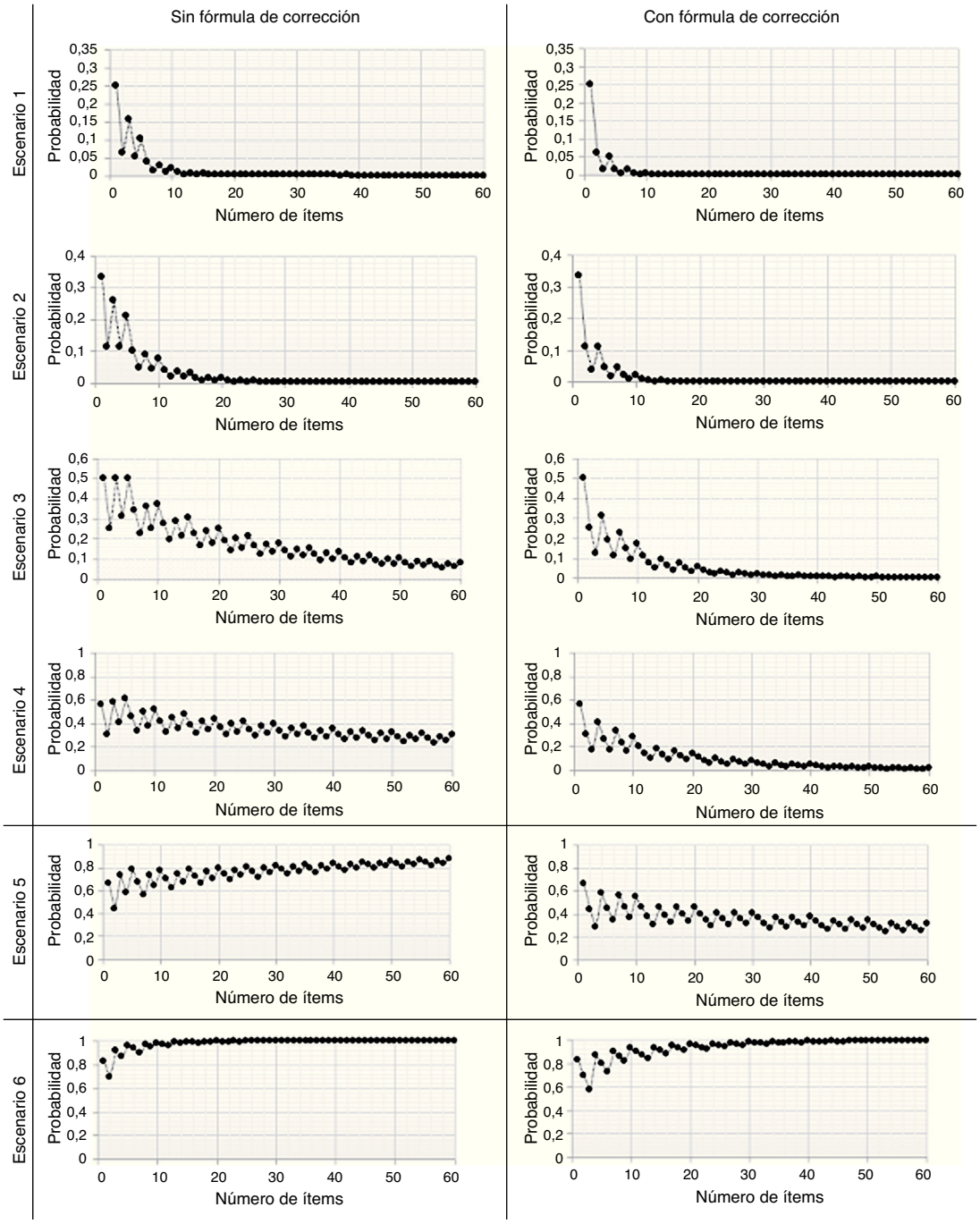

Figura 2 Probabilidad, bajo el modelo Nedelsky, de aprobar un examen en función del número de ítems, con y sin fórmula de corrección, para cada escenario descrito en la tabla 1.

distractores constituyen un límite natural y asequible para un IOM, y se sugiere que sólo se incluyan más si se logra asegurar la calidad y pertinencia de los excedentes[41–43]. Además, existe suficiente evidencia sobre exámenes con tres o cuatro distractores en donde al menos uno de ellos es no funcional[6,10,11,14,39,44,45], lo cual propicia un aumento del efecto de EG y TW y hace debatibles los juicios que se basan en los resultados del examen, principalmente para puntajes cercanos al estándar de pase y preocupantemente si se trata de un examen de altas consecuencias. Desde el contexto de educación médica, es evidente la preocupación asociada con acertar por adivinación, especialmente cuando se asume correspondencia entre el porcentaje de respuestas correctas y el porcentaje de preguntas en las que el estudiante tiene el conocimiento suficiente para reconocer la respuesta correcta.

La fórmula de corrección tradicional, que resta $1/(m-1)$ para cada respuesta incorrecta, resulta sustancialmente conveniente para estudiantes de nivel medio (Escenarios 3, 4 y 5), donde es imperativo discriminar entre los que dominan con suficiencia el tema y aquellos que no alcanzan la competencia mínima. Empero, también puede afectar a

estudiantes aversos al riesgo y no solo reducir su puntuación, sino mermar su probabilidad de aprobar[22]. En efecto, adivinar casi siempre es ventajoso, incluso con la fórmula de corrección tradicional, porque es infrecuente que el sustentante sea incapaz de, al menos, descartar un distractor[4,46]. Existen dos circunstancias propiciadas por la implementación de la fórmula que contradicen su objetivo: *a)* la fuerte influencia de ésta hacia el efecto la TA (más la consecuente carga de varianza irrelevante al constructo)[22,24], y *b)* su utilización para paliar la validez perdida en el contenido[15]. Como alternativas a la fórmula de corrección tradicional, se ha sugerido aumentar la penalización por ítem incorrecto para desalentar a los sustentantes que ignoran el discurso precautorio[47], o modificar el estándar de pase en función al desempeño de los sustentantes y a la probabilidad de adivinar[48].

Los Escenarios 4, 5 y 6 parecen más plausibles que simplemente asumir que los ítems son uniformes (es decir, que en todos sucede lo mismo: adivinar a ciegas, o descartar determinado número de distractores). En el ámbito de educación médica, especialmente para exámenes de medianas y altas consecuencias, es totalmente verosímil asumir que los estudiantes poseen un cierto nivel de θ y de TW debido a su trayectoria académica y experiencia con EOM, por eso resultan más apropiados los supuestos en los que la asignación de probabilidad varía. Es importante enfatizar que en el análisis teórico se supone que, sin importar la valoración que asigne el individuo a cada opción o su concepción respecto a adivinar, siempre da respuesta a los ítems. En realidad, la decisión sobre responder los ítems introduce otra fuente de incertidumbre que no se investigó en este análisis.

Con base en los resultados de este trabajo, es posible recomendarle a los tomadores de decisiones que consideren los elementos ajenos al rasgo que se pretende medir que influyen en el puntaje obtenido en el examen, o al menos estén conscientes de ellos, de modo que el punto de corte no sea arbitrariamente localizado en un porcentaje de la calificación máxima, y/o que se establezcan estándares para asegurar la calidad de los reactivos que constituyen las pruebas. Finalmente, como futura línea de investigación, sugerimos explorar experimentalmente la pertinencia del modelo teórico propuesto con la intención principal de determinar el efecto de la regla de calificar sobre los factores TA, TW y EG. En específico, podría definirse un método para evaluar cómo la regla para calificar modifica el efecto de la TA, a través de la percepción del riesgo, utilizando ítems que difieran en el grado que permiten la ocurrencia de TW y EG.

Contribución de cada autor

AJN concibió múltiples ideas, desarrolló el trabajo y llevó a cabo los análisis teóricos basados en el modelo Nedelsky. IL concibió las ideas principales del artículo, determinó la estructura de las ideas, y desarrolló el programa informático para el análisis teórico. Ambos autores aprobaron la versión final del artículo.

Presentaciones previas

Ninguna.

Responsabilidades éticas

Protección de personas y animales. Los autores declaran que para esta investigación no se han realizado experimentos en seres humanos ni en animales.

Confidencialidad de los datos. Los autores declaran que en este artículo no aparecen datos de pacientes.

Derecho a la privacidad y consentimiento informado. Los autores declaran que en este artículo no aparecen datos de pacientes.

Financiación

Ninguna.

Conflictos de interés

Los autores declaran no tener conflictos de interés.

Agradecimientos

Ninguno

Referencias

1. Downing SM. Assessment of knowledge with written test forms. En: Norman GR, Van der Vleuten C, Newble DI, editores. International handbook of research in medical education Volume II. Dorcrecht: Kluwer Academic Publishers; 2002. p. 647–72.
2. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. Med Teach. 2004;26(8):709–12.
3. Betts LR, Elder TJ, Hartley J, Trueman M. Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes? Assess Eval High Educ. 2009;34(1):1–15.
4. Chang SH, Lin PC, Lin ZC. Measures of partial knowledge and unexpected responses in multiple-choice tests. J Educ Techno Soc. 2007;10(4):95–109.
5. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple choice item-writing guidelines for classroom assessment. Appl Meas Educ. 2002;15(3):309–34.
6. Jozefowicz RF, Koeppen BM, Case S, Galbrath R, Swanson D, Glew RH. The quality of in-house medical school examinations. Acad Med. 2002;77(2):156–61.
7. Ware J, Torstein V. Quality assurance of item-writing: during the introduction of multiple choice questions in medicine for high stakes examinations. Med Teach. 2009;31(3):238–43.
8. Millman J, Bishop C, Ebel R. An analysis of test-wiseness. Educ Psychol Meas. 1965;25:707–26.
9. Downing SM. Guessing on selected-response examinations. Med Educ. 2003;37(8):670–1.
10. Rogausch A, Hofer R, Krebs R. Rarely selected distractors in high stakes medical multiple- choice examinations and their recognition by item authors: a simulation and survey. BMC Med Educ. 2010;10:85.
11. Jurado-Núñez A, Flores-Hernández F, Delgado-Maldonado L, Sommer-Cervantes H, Martínez-González A, Sánchez-Mendiola M. Distractores en preguntas de opción múltiple para estudiantes de medicina: ¿cuál es su comportamiento en un examen sumativo de altas consecuencias? Inv Ed Med. 2013;2(8):202–10.

12. Chiu, TW. Correction for guessing in the framework of the 3PL item response theory. Disertación doctoral, 2011 [consultado 2 Dic 2014]. Disponible en: https://rucore.libraries.rutgers.edu/rutgers-lib/27294/pdf/1
13. Leenen I. Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. Inv Ed Med. 2014;3(9):40–55.
14. Downing SM. Construct-irrelevant variance and flawed test questions: do multiple choice item-writing principles make any difference? Acad Med. 2002;77(10):S103–4.
15. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol. 1995;50(9):741–9.
16. Downing SM, Reliability:. on the reproducibility of assessment data. Med Educ. 2004;38(9):1006–12.
17. Gempp R. El error estándar de medida y la puntuación verdadera de los tests psicológicos: algunas recomendaciones prácticas. Ter Psicol. 2006;24(2):117–30.
18. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37(9):830–7.
19. Muñiz J. La validez desde una óptica psicométrica. Acta Comport. 2005;13(1):9–20.
20. Burton R:. Multiple-choice and true/false tests: myths and misapprehensions. Assess Eval High Educ. 2005;30(1):65–72.
21. Suh Y, Bolt D. Nested logit models for multiple-choice item response data. Psychometrika. 2010;75(3):454–73.
22. Lesage E, Valcke M, Sabbe E. Scoring methods for multiple-choice assessment in higher education –Is it still a matter of number right scoring or negative marking? Stud Educ Eval. 2013;39(3):188–93.
23. Burton R. Multiple choice and true/false tests: reliability measures and some implications of negative marking. Assess Eval High Educ. 2004;29(5):585–95.
24. Bar-Hillel M, Budescu D, Attali Y. Scoring and keying multiple choice tests: a case study in irrationality. Mind Soc. 2005;4:3–12.
25. Ziller R. A measure of the gambling response-set in objective tests. Psychometrika. 1957;22(3):289–92.
26. Pérez-Gil, JA. Modelos de Medición: Desarrollos actuales, supuestos, ventajas e inconvenientes: Teoría de Respuesta a los Items (TRI). Apuntes de la asignatura: Desarrollos actuales de la medición: Aplicaciones en evaluación psicológica (Tema 1). Departamento de Psicología Experimental. Universidad de Sevilla. [consultado 2 Dic 2014]. Disponible en: http://innoevalua.us.es/files/irt.pdf
27. Muñiz J. Las teorías de los tests: Teoría Clásica y Teoría de Respuesta a los Ítems. Papeles Psicol. 2010;31(1):57–66.
28. Hambleton R, Jones R. Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. Educ Meas. 1993;12(3):38–47.
29. Erguven M. Two approaches to psychometric process: Classical test theory and item response theory. IBSU J Educ. 2014;2(2):23–30. Disponible en: http://journal.ibsu.edu.ge/index.php/sje/article/view/537
30. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1980 (Trabajo original publicado en 1960).
31. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. En: Lord FM, Novick MN, editores. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968. p. 396–479.
32. Cao J, Stokes L, Bayesian IRT. guessing models for partial guessing behaviors. Psychometrika. 2008;73(2):209–30.
33. San Martín E, Del Pino G, De Boeck P. IRT Models for Ability-Based Guessing. Appl Psychol Meas. 2006;30(3):183–203.
34. Thissen D, Steinberg L. A response model for multiple choice items. Psychometrika. 1984;49(4):501–19.
35. Levine M, Drasgow F. The relation between incorrect option choice and estimated ability. Educ Psychol Meas. 1983;43:675–85.
36. Bock D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika. 1972;37(1):29–51.
37. Samejima F. A new family of models for the multiple-choice item. Knoxville: University of Tennessee, Department of Psychology; 1979.
38. Bechger T, Verstralen H, Maris G, Verhelst N. The Nedelsky model for multiple choice items. En: van der Ark LA, Croon MA, Sijtsma K, editores. New developments in categorical data analysis for the social and behavioral sciences. Mahwah, NJ: Lawrence Erlbaum Associates; 2005. p. 187–206.
39. Downing SM. The effects of violating standard item writing principles on tests and students: The consequences of using flawed items on achievement examinations in medical education. Adv Health Sci Educ. 2005;10(2):133–43.
40. Tarrant M, Knierim A, Hayes S, Ware J. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Educ Pract. 2006;6(6):354–63.
41. Cizek GJ, O'Day DM. Further investigation of nonfunctioning options in multiple-choice test items. Educ Psychol Meas. 1994;54(4):861–87.
42. Abad FJ, Olea J, Ponsoda V. Analysis of the optimum number alternatives from the Item Response Theory. Psicothema. 2001;13(1):152–8.
43. Rodriguez MC. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educ Meas Issues Pract. 2005;24(2):3–13.
44. Cizek GJ, Robinson L, O'Day DM. Nonfunctioning options: A closer look. Educ Psychol Meas. 1998;58(4):605–11.
45. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non- functioning distractors in multiple-choice questions: a descriptive analysis. BMC Med Educ. 2009;9:40.
46. MacCann R. Reliability as a function of the number of item options derived from the ''knowledge of random guessing'' model. Psychometrika. 2004;69(1):147–57.
47. Espinosa MP, Gardeazabal J. Optimal correction for guessing in multiple-choice tests. J Math Psychol. 2010;54(5):415–25.
48. Dochy F, Kyndt E, Baeten M, Pottier S, Veestraeten M. The effects of different standard setting methods and the composition of borderline groups: A study within a law curriculum. Stud Educ Eval. 2009;35(4):174–82.