Amenazas a la validez en evaluación: implicaciones en educación médica

Blanca Ariadna Carrillo Avalos[a,*], Melchor Sánchez Mendiola[b], Iwin Leenen[c]

Facultad de Medicina

Resumen

Las amenazas a la validez en evaluación educativa son elementos que interfieren con la interpretación propuesta de los resultados de una prueba, pueden ocurrir tanto en exámenes escritos como en pruebas de desempeño y evaluación de competencias clínicas. Estas amenazas se suelen agrupar en dos clases principales: subrepresentación del constructo y varianza irrelevante al constructo. La primera se refiere a que en la prueba no haya suficientes ítems, casos u observaciones para generalizar apropiadamente al dominio completo que se pretende evaluar. La segunda tiene que ver con la presencia de sesgos que interfieren de manera sistemática con la interpretación de los resultados de una prueba, como pueden ser la calidad de los ítems y errores sistemáticos de los evaluadores, entre otros factores que pueden influir sobre la puntuación obtenida. En este artículo se describen las características de las amenazas principales, su importancia y algunas recomendaciones para evitarlas al elaborar y aplicar instrumentos de evaluación en ciencias de la salud. La comprensión de estas amenazas es útil para desarrollar pruebas cuyos resultados tengan niveles aceptables de validez que nos permitan conocer mejor el desempeño de los estudiantes.

Palabras clave: *Amenazas a la validez; evaluación del aprendizaje; educación médica; México.*

[a]Departamento de Ciencias Morfológicas, Facultad de Medicina, Universidad Autónoma de San Luis Potosí, S. L. P., México.
[b]División de Estudios de Posgrado, Facultad de Medicina, Universidad Nacional Autónoma de México, Cd. Mx., México.
[c]División de Estudios de Posgrado, Facultad de Psicología, Universidad Nacional Autónoma de México, Cd. Mx., México.
Recibido: 10-diciembre-2019. Aceptado: 17-febrero-2020.
*Autora para correspondencia: Blanca Ariadna Carrillo Avalos. Av. Venustiano Carranza 2405, Col. Los Filtros, San Luis Potosí, San Luis Potosí, México. CP 78210. Teléfono: 4448 2623 45, ext.: 6635. Correo electrónico: bariadna@gmail.com
La revisión por pares es responsabilidad de la Universidad Nacional Autónoma de México.

https://doi.org/10.22201/facmed.20075057e.2020.34.221

Threats to validity in assessment: implications in medical education

Abstract

Validity threats in educational assessment are elements that interfere with the proposed interpretation of a test score. They can occur in written tests as well as in performance and clinical competency assessments. They are usually grouped in two major categories: construct underrepresentation and construct-irrelevant variance. The former refers to tests with insufficient items, cases, or observations to make a proper generalization towards the full to-be-assessed domain. The latter is related to the presence of biases that can interfere systematically with the interpretation of a test score, such as item quality and raters' systematic errors, among other factors that may have an effect on the obtained score. In this paper we describe the characteristics of some of these threats, their importance, and some recommendations to avoid them during the development of assessment instruments in health sciences education. The insights offered can be useful to devise tests and assessment instruments that allow us to draw more valid inferences about students' knowledge and abilities.

Keywords: *Validity; validity threats; learning assessment; medical education; Mexico.*

INTRODUCCIÓN

El análisis de la validez de los usos e interpretaciones de las puntuaciones de una prueba nos informará sobre el grado en que son apropiados estos usos e interpretaciones para los fines de la evaluación. Sin embargo, la tarea de validación no termina aquí, ya que es necesario descartar otras hipótesis que puedan explicar resultados que no concuerden con la hipótesis original, e identificar elementos que puedan interferir con la interpretación apropiada de los resultados[1-3]. Estas hipótesis apuntan a posibles amenazas a la validez y considerarlas dará mayor fortaleza a las decisiones que se tomen con base en las puntuaciones del examen que estamos valorando. Este análisis cobra mayor relevancia mientras mayor sea el escrutinio al que esté sometido el proceso de evaluación, y mayores sean las potenciales consecuencias del uso de los resultados en los sustentantes, los docentes y las instituciones educativas.

En otro artículo revisamos el concepto moderno de validez en evaluación educativa y su relevancia en educación médica[4]. En este trabajo describiremos las principales amenazas a la validez que existen en evaluación educativa, sus implicaciones en educación en ciencias de la salud y algunas recomendaciones para evitarlas.

Las amenazas a la validez son factores que interfieren con la interpretación del significado de la puntuación obtenida en la evaluación[2,3]. Pueden encontrarse en cualquier tipo de evaluación, ya sea de conocimientos teóricos o prácticos, diagnóstica, formativa o sumativa[3]. En muchas ocasiones los exámenes que se aplican en las escuelas y facultades de medicina, enfermería y otras ciencias de la salud se hacen por medio de preguntas de opción múltiple (POM)[5,6], en este artículo nos enfocaremos principalmente en este tipo de pruebas, aunque las amenazas a la validez se pueden presentar –y deben considerarse– también en evaluaciones prácticas como el examen clínico objetivo estructurado (ECOE). Con respecto a las evaluaciones con POM, se han publicado varios estudios que documentan que la calidad de los reactivos o ítems es limitada[7-9], ya que con frecuencia no se elaboran con el profesionalismo necesario ni siguiendo los lineamientos técnicos para ello[6].

Aunque se mencionan varios tipos de amenazas (por ejemplo, Crooks, Kane y Cohen consideran al menos 23, relacionadas con ocho inferencias)[10], en general se agrupan en dos clases principales: la subrepresentación del constructo (SC) y la varianza irrelevante al constructo (VIC)[11]. A continuación explicamos estos dos conceptos.

Según la teoría clásica de los test (TCT), la pun-

tuación observada *(X)* es una combinación de la puntuación verdadera *(true = T)*, más un componente de error aleatorio *(random error = E_r)*:[12,13]

$$X = T + E_r.$$

En esta fórmula, la puntuación verdadera *T* resulta de todos los factores que tienen un efecto sistemático sobre la puntuación observada *X*, incluyendo tanto el constructo de interés como otros factores sistemáticos que no son el objetivo de la medición (por ejemplo, gran severidad de un examinador en un ECOE que cause disminución sistemática de las puntuaciones). Por otro lado, el error aleatorio (E_r) recoge el efecto de todas las circunstancias que afectan la puntuación observada de manera no sistemática, es decir factores que varían cada vez que se aplica la prueba, como el cansancio o estrés del alumno[14]. Tanto la puntuación verdadera como el error aleatorio son constructos hipotéticos y desconocidos, pero por medio de métodos de la TCT se pueden hacer conclusiones a partir de una muestra[15].

La discusión anterior indica que la puntuación verdadera puede descomponerse en dos partes: la puntuación en el constructo de interés (ɵ) más la puntuación que se debe a otros factores sistemáticos. Como la segunda parte incluye efectos de factores no intencionados, Haladyna y Downing[14] la denominan el error sistemático *(E_s)* y obtienen la siguiente fórmula:

$$X = \text{ɵ} + E_s + E_r. \quad (1)$$

A partir de esta fórmula, se definen los conceptos de SC y VIC. Por un lado, existe una amenaza a la validez cuando la medición de ɵ es a través de ítems que no son representativos del dominio completo a evaluar; es decir, cuando los ítems de la prueba evalúan *de manera incompleta* el constructo que se desea medir. Este caso se considera SC. Por otro lado, la VIC está asociada con el error sistemático E_s, el cual es causado por la medición involuntaria de constructos irrelevantes –cuya medición no es el objetivo del examen–, por lo que interfieren con la medición del constructo original y por lo tanto con la validez de la interpretación de la puntuación[2,11,14].

Mención aparte merece el componente E_r de la fórmula (1). Por definición, este componente no produce SC ni VIC, ya que su efecto no es sistemático. Sin embargo, la varianza debido a E_r no es deseable y también constituye una amenaza a la validez. En el marco de la TCT, los factores reunidos en E_r conllevan una baja confiabilidad (y un error estándar de medición grande)[2,9,16]. En este sentido, la fórmula (1) permite ilustrar la diferencia entre validez y confiabilidad. Por un lado, tanto E_r y E_s se refieren a errores a la medición del constructo y, por lo tanto, ambos constituyen amenazas a la validez; por otro lado, solo E_r causa varianza no sistemática y, por lo tanto, solo este factor está asociado con la (baja) confiabilidad. Esto aclara por qué confiabilidad se considera un prerrequisito para validez. En el resto de este artículo solo se considerarán amenazas a la validez relacionadas con factores sistemáticos: SC y VIC.

SUBREPRESENTACIÓN DEL CONSTRUCTO (SC)

En el caso de una prueba escrita, la SC se refiere a que, considerando el universo de ítems o preguntas posibles relevantes al dominio explorado, la prueba esté integrada por una muestra de ítems que puede:

- Tener muy pocos ítems y ser insuficiente para evaluar el dominio del conocimiento correspondiente,
- Estar sesgada hacia un área del tema a evaluar, convirtiéndose en una muestra no representativa,
- Evaluar contenido trivial o factual al nivel más bajo de la pirámide de Miller[2,9,17].

La SC es una amenaza particularmente importante para la inferencia de extrapolación, ya que la interpretación de las puntuaciones es más limitada si los resultados no son representativos del constructo que se supone que la prueba evalúa[18].

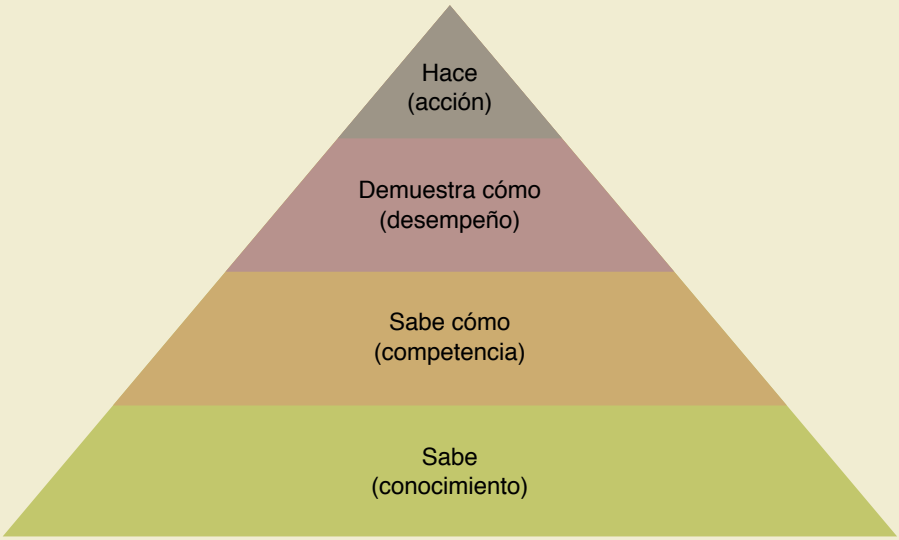
Utilizaremos para ilustrar las distintas amenazas a la validez un ejemplo de ciencias básicas: el tema de anatomía de la cabeza. Este tema, sin neuroanatomía, abarca 160 páginas del libro de "Anatomía con orientación clínica de Moore"[19], uno de los libros más utilizados para la enseñanza de anatomía humana en México. Si aplicáramos el examen de la **tabla 1** con el objetivo de evaluar los conocimientos de anatomía representados en el libro de Moore, las amenazas a la validez con respecto a la SC serían las siguientes:

Tabla 1. Ejemplos de preguntas de un examen de anatomía de la cabeza

Pregunta	Opciones de respuesta
1. ¿Cuántos huesos conforman el viscerocráneo?	a. 11 b. 12 c. 13 d. 14 e. 15*
2. La siguiente estructura generalmente está inervada por el nervio laríngeo interno:	a. Aritenoides oblicuo b. Cricoaritenoideo posterior c. Cricotiroideo d. Mucosa infralaríngea e. Mucosa supralaríngea*
3. En la coroides no ocurre lo siguiente:	a. Contiene ramas de la arteria central de la retina* b. La lámina coroidocapilar es la más interna c. Produce el reflejo rojo del fondo de ojo d. Se encuentra entre la esclera y la retina e. Sus venas drenan en una vena vorticosa
4. Una mujer joven se golpea la cabeza con el cuadro de mandos del automóvil durante una colisión frontal. A continuación, sufre un desgarro de la parte frontal del cuero cabelludo con sangrado abundante. La herida se lava con suero fisiológico y se cubre con una venda estéril. Cuando la mujer llega al hospital tiene los dos ojos morados. En la exploración posterior no se aprecia ninguna lesión ocular[19]. ¿Cuál es la arteria que más probablemente se lesionó en este caso?	a. Auricular posterior b. Facial, porción cervical c. Mentoniana d. Supraorbitaria* e. Temporal superficial
5. ¿Cuál es la acción principal del músculo recto inferior? I. Abducir el globo ocular II. Aducir el globo ocular II. Descender el globo ocular IV. Rotar lateralmente el globo ocular V. Rotar medialmente el globo ocular	a. I, II y III b. II, III y IV* c. III, IV y V d. I, III y V e. I y IV
6. Which bone does NOT contribute to the orbit?	a. Frontal bone b. Maxilla c. Palate bone d. Sphenoid bone e. Temporal bone*
7. Un boxeador recibió un golpe en la cara lateral de la nariz, quedando deformada y con los huesos nasales desplazados. Asimismo, presentaba una rotura de los cartílagos de la nariz, epistaxis y obstrucción de la vía respiratoria nasal. ¿Cuál es la arteria en donde se origina la epistaxis?	a. Etmoidal anterior b. Nasal lateral* c. Supraorbitaria d. Supratroclear e. Transversa de la cara
8. ¿Cuál de los siguientes es un músculo de la cara?	a. Bíceps braquial b. Dorsal ancho c. Esternocleidomastoideo d. Frontal* e. Psoas mayor
* Respuesta correcta	

- *Número de preguntas insuficiente.* Un examen que consta de 8 ítems no será adecuado a la luz del amplio universo de ítems de anatomía de la cabeza que se pueden considerar, con base en la extensión de los temas que comprende esta unidad y los objetivos de aprendizaje que se hayan establecido en el currículo. Downing y Haladyna[2] sugieren un mínimo de 30 preguntas en general, mientras que en el manual del *National Board of Medical Examiners* sugieren 100 preguntas para obtener resultados reproducibles[20], aunque estos autores no especifican el tipo de prueba al que van dirigidas estas recomendaciones. En general, para determinar la cantidad adecuada de ítems

Figura 1. Pirámide de Miller[17]



se sugiere considerar los resultados de aprendizaje establecidos en la tabla de especificaciones y factores como el tiempo real que tienen los alumnos para contestar el examen, así como ponderar la importancia de cada uno de los temas a examinar. También es relevante si la prueba es una evaluación sumativa o formativa, así como la exactitud necesaria de las puntuaciones[21,22].

- *Sesgo.* Esta amenaza puede presentarse en caso de que los ítems solo examinen una parte de los temas establecidos en la tabla de especificaciones de la evaluación, sin incluir otras porciones importantes de dicha tabla[2,9].
- *Nivel de evaluación con base en la pirámide de Miller.* Un marco de referencia utilizado en educación médica es la pirámide de Miller (**figura 1**), en la que se proponen los niveles de desarrollo académico y profesional a evaluar, así como una estructura de evaluación y planeación de actividades de aprendizaje[23,24]. Esta amenaza se refiere a que, en el caso de que los objetivos de aprendizaje y evaluación contemplaran niveles de competencia, desempeño o ejecución en la pirámide de Miller, las preguntas fueran mayoritariamente acerca de hechos memorizables (como las preguntas 1, 6 y 8 de la **tabla 1**), y que no evaluaran niveles superiores como la integración entre estos conocimientos y otros previamente adquiridos, ni su relación con la aplicación clínica o con los contenidos de otras asignaturas cuyos temas estén relacionados con las estructuras estudiadas. En una ciencia básica como anatomía, no es fácil elaborar ítems que vayan más allá de conocimientos factuales; sin embargo, es posible lograrlo mientras se tengan claros los objetivos de aprendizaje y los de evaluación en la tabla de especificaciones, así como los usos e interpretaciones de los resultados de la prueba[25].

VARIANZA IRRELEVANTE AL CONSTRUCTO (VIC)

Como ya se mencionó, la VIC se origina del error sistemático debido a una variable irrelevante al constructo que se pretende medir[14]. A continuación, discutimos algunas características de un examen que suelen ocasionar VIC y las ilustramos con el mismo ejemplo del examen de 8 preguntas en la **tabla 1**:

- *Ítems mal elaborados.* Es importante conocer las características de una POM de calidad, descritas en varios documentos[20,26], para evitar ítems defectuosos que puedan causar mayor dificultad para contestarlos o que incluso presenten pistas basadas en aspectos formales para determinar la respuesta correcta[27]. Por ejemplo, en la pregunta 2 de la **tabla 1** no sabemos qué significa "generalmente" ni a qué tipo de estructuras se refiere (¿músculos?). Además, la respuesta correcta es la única estructura que parece no ser un músculo. Otros defectos consisten en elaborar preguntas con opciones que incluyan "todas las anteriores" o "ninguna de las anteriores"[28]. Otro tema ampliamente estudiado en la elaboración de POM es la cantidad de opciones[29-31].
- *Lenguaje.* Por su nivel, dificultades o ambigüedad en la redacción, los formatos muy complicados o extensos (como la pregunta 4 de la **tabla 1**), hacen que el sustentante pase más tiempo leyendo que determinando la respuesta correcta, y esto debe considerarse con respecto al tiempo real que se tiene para presentar la prueba[20]. Un defecto común es elaborar preguntas que pueden confundir al alumno; es decir, preguntas que, aunque sí conoce la respuesta, podría contestar mal: por ejemplo, al contestar la pregunta 5 de la **tabla 1**, el alumno que podría saber las funciones del músculo referido, tendrá que pasar tiempo relacionando los números romanos con la opción correcta. Además, primero debe saber los números romanos[20]. La estructura de las oraciones debe ser lo suficientemente clara y evitar el uso de jerga para que no sea causa de respuestas equivocadas. Un ejemplo es la pregunta 7 de la **tabla 1**, que contiene la palabra "epistaxis" que puede ser confusa para un estudiante de primer año, pues todavía no conoce los términos clínicos[32].
- *Formato en negativo.* Chiavaroli[33] explica que deben evitarse preguntas que incluyen negaciones como las preguntas 3 y 6 de la **tabla 1**. Esto es debido a que existe un doble negativo (en el sentido de que la pregunta incluye una negación e identificar las opciones incorrectas implica negarlas –decir que *no* son correctas–), por lo que existe el riesgo de que el alumno no identifique la parte negativa de la pregunta (aunque la palabra "excepto" o "no" se encuentre en negritas), y que la forma de contestar no se lleva a cabo mediante el proceso de respuesta deseado, afectando así esta evidencia de validez. En el caso de POM en asignaturas de ciencias clínicas, se puede evitar la negación utilizando términos como "cuál es la contraindicación o el riesgo".
- *Funcionamiento diferencial de ítem* (*differential item functioning*; *DIF).* El DIF significa que los sustentantes con características o antecedentes distintos (por ejemplo, de diferente sexo o nivel socioeconómico) no tienen la misma probabilidad de responder de manera correcta *a pesar de poseer el mismo nivel en el constructo que se desea medir.* Además de diferencias de género o nivel socioeconómico, un análisis DIF puede comparar grupos diferentes con respecto a características demográficas, religiosas, culturales o lingüísticas[34]. Un ejemplo de esta amenaza se presenta en la pregunta 6 de la **tabla 1**: está en otro idioma (además de que su formato es negativo), de tal manera que los alumnos que no sepan inglés, aunque tengan el conocimiento evaluado por esta pregunta (y suponiendo que el inglés no es parte del constructo que se desea medir), podrían responder incorrectamente[35]. Otro ejemplo es cuando preguntamos las manifestaciones de la lesión del nervio mediano con el término "mano de predicador"; los estudiantes de algunas religiones pueden no entender a qué se refiere.
- *Discordancia con el dominio.* Si entre los objetivos de aprendizaje no se establece el estudio de un tema en particular, sería equivocado evaluarlo, ya que esto causaría que los ítems no correspondieran al dominio de contenido que se pretende evaluar[2].
- *Hacer trampa.* Hay muchas formas de hacer trampa en los exámenes: copiar al compañero de junto, usar un "acordeón" o algo similar, tener acceso a las preguntas de manera previa a la presentación del examen, y hasta el uso de los *smart watches*[36]. Estos comportamientos pueden generar falsos positivos y en este sentido introducir varianza sistemática no deseada en las puntuaciones de las pruebas[9].
- *Enseñar a la prueba ("teaching to the test").* Se refiere a que los alumnos reciban entrenamiento

para contestar los ítems de una prueba en particular, incluso practicando con las preguntas que aparecerán en el examen real. Esta práctica es una amenaza para la validez porque los alumnos están aprendiendo las respuestas de memoria sin adquirir el conocimiento que están evaluando las preguntas; de esta manera no es posible generalizar el resultado hacia el resto del universo de ítems posibles que evalúan el constructo deseado[37].

- *Testwiseness*. Con base en la gran cantidad de exámenes de opción múltiple que contestan durante su vida académica, se considera que muchos estudiantes de medicina son *test wise*[28]. Quiere decir que han desarrollado estrategias para contestar exámenes deduciendo cuál es la respuesta correcta con base en la estructura gramatical y de redacción: opciones más largas, opciones con más detalles, etc. El dominio de dichas estrategias es irrelevante al constructo, ya que causa que las respuestas no reflejen lo que los estudiantes saben realmente[38]. Es importante distinguir el *testwiseness* de otros conceptos como el *educated guessing*[39], de manera que con el primero se pueden conseguir respuestas correctas, aun sin tener conocimiento; mientras que con el segundo los alumnos logran eliminar opciones con base en el rasgo latente que se desea medir por medio de la evaluación, pero no consiguen identificar por completo la respuesta correcta, por lo que terminan adivinando.

CONCLUSIONES

Las amenazas a la validez resultan aspectos importantes a tomar en cuenta durante la planeación y desarrollo de una prueba, ya que su presencia disminuye la validez de sus resultados, confunde la interpretación propuesta de los mismos y lleva a conclusiones e inferencias erróneas.

Cuando planeamos y desarrollamos pruebas para evaluar eficazmente el constructo deseado, es necesario que capacitemos y motivemos a los elaboradores de preguntas de nuestras escuelas para que tengan "la voluntad de invertir bastante tiempo y esfuerzo en crear preguntas de opción múltiple efectivas"[9]. Tomar en cuenta las amenazas a validez descritas permite afrontarlas y corregirlas antes de que ocurran y afecten las interpretaciones de las puntuaciones de la prueba. Debemos adoptar una actitud más proactiva hacia la prevención de estas amenazas, incluyendo su descripción y efectos en las actividades de formación docente.

Con respecto a las amenazas por subrepresentación del constructo, una recomendación fundamental es establecer claramente, desde la tabla de especificaciones, los objetivos de aprendizaje y el dominio explorado, así como la importancia y la proporción de preguntas que deberán asociarse a cada subtema. Por otro lado, la varianza irrelevante de constructo puede disminuirse significativamente al desarrollar habilidades para la elaboración correcta de ítems de opción múltiple.

Debemos impartir talleres de elaboración de preguntas, tanto para ciencias básicas como para ciencias clínicas; un comité evaluador con experiencia en la elaboración correcta de preguntas debe revisar de forma colegiada el instrumento de evaluación antes y después de su aplicación. Asimismo, sería recomendable incluir en la prueba preguntas que consideren varios niveles de la pirámide de Miller, para ampliar y profundizar el abanico de evaluación de los profesionales de la salud.

REFERENCIAS

1. Cronbach LJ. Five perspectives on validity argument. En: Wainer H, Braun HI, editores. Test validity [Internet]. New York: Routledge; 1988. p. 3-17. Disponible en: https://doi.org/10.4324/9780203056905
2. Downing SM, Haladyna TM. Validity threats: Overcoming interference with proposed interpretations of assessment data. Med Educ. 2004;38(3):327-33.
3. Downing SM, Yudkowski R, editores. Assessment in health professions education. New York and London: Routledge; 2009. 317 p.
4. Carrillo BA, Sánchez M, Leenen I. El concepto moderno de validez y su uso en educación médica. Inv Ed Med. 2020; 9(33):98-106.
5. Norman G, van der Vleuten C, Newble D. International Handbook of Research in Medical Education. Norman G, van der Vleuten C, Newble D, editores. Springer; 2002. 1106 p.
6. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. Acad Med. 2002;77(2):156-61.
7. Ware J, Vik T. 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. Tarrant M, Knierim A, Hayes SK, Ware J. The frequency of item writing flaws in multiple-choice questions used in

high stakes nursing assessments. Nurse Educ Today. 2006; 26(8):662-71.
9. Downing SM. Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. Adv Heal Sci Educ. 2002;7(3):235-41.
10. Crooks TJ, Kane MT, Cohen AS. Threats to the valid use of assessments. Assess Educ Princ Policy Pract. 1996;3(3):265-85.
11. Messick S. Validity. En: Linn RL, editor. Educational Measurement [Internet]. New York: Macmillan; 1989. p. 13-103. Disponible en: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2330-8516.1987.tb00244.x
12. Schuwirth LWT, Van Der Vleuten CPM. General overview of the theories used in assessment: AMEE Guide No. 57. Med Teach. 2011;33(10):783-97.
13. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ. 2010;44(1):109-17.
14. Haladyna TM, Downing SM. Construct-Irrelevant Variance in High-Stakes Testing. Educ Meas Issues Pract [Internet]. 2004;23(1):17-27. Disponible en: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3992.2004.tb00149.x
15. 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.
16. Downing SM. Reliability : on the reproducibility of assessment data. Med Educ. 2004;38:1006-12.
17. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(9):S63-7.
18. Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. Constructing a validity argument for the mini-clinical evaluation exercise: A review of the research. Acad Med. 2010;85(9):1453-61.
19. Moore K, Dailey A, Agur A. Anatomía con orientación clínica. 7a ed. Philadelphia: Wolters Kluwer Health, S.A., Lippincot Williams & Wilkins; 2013.
20. National Board of Medical Examiners. Cómo elaborar preguntas para evaluaciones escritas en el área de ciencias básicas y clínicas. 4th ed. Paniagua MA, Swygert KA, editores. Philadelphia, PA: National Board of Medical Examiners; 2016. 100 p.
21. Moreno R, Martínez RJ, Muñiz J. Directrices para la construcción de ítems de elección múltiple. Psicothema [Internet]. 2004;16(3):490-7. Disponible en: https://www.redalyc.org/articulo.oa?id=72716324
22. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. STANDARDS for Educational and Psychological Testing. 6th ed. American Educational Research Association. Washington, D. C.: American Educational Research Association, American Psychological Association & National Council on Measurement in Education; 2014. 243 p.
23. Williams BW, Byrne PD, Welindt D, Williams M V. Miller's pyramid and core competency assessment: A study in relationship construct validity. J Contin Educ Health Prof. 2016;36(4):295-9.
24. Pangaro L, Ten Cate O. Frameworks for learner assessment in medicine: AMEE Guide No. 78. Med Teach. 2013;35:e1197-e1210.
25. Hadie SNH. The Application of Learning Taxonomy in Anatomy Assessment in Medical School. Educ Med J. 2018;10(1):13-23.
26. 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.
27. Downing SM. Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? Acad Med. 2002;77(10 SUPPL.):103-4.
28. Downing SM. The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. Adv Heal Sci Educ. 2005;10(2):133-43.
29. Abad FJ, Olea J, Ponsoda V. Analysis of the optimum number alternatives from the Item Response Theory. Psicothema. 2001;13(1):152-8.
30. 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.
31. Haladyna TM, Rodriguez MC, Stevens C. Are Multiple-choice Items Too Fat? Appl Meas Educ [Internet]. 2019;32(4):350-64. Disponible en: https://doi.org/10.1080/08957347.2019.1660348
32. Hicks NA. Guidelines for identifying and revising culturally biased multiple-choice nursing examination items. Nurse Educ. 2011;36(6):266-70.
33. Chiavaroli N. Negatively-worded multiple choice questions: An avoidable threat to validity. Pract Assessment, Res Eval. 2017;22(3):1-14.
34. Gómez-Benito J, Sireci S, Padilla JL, Dolores Hidalgo M, Benítez I. Differential item functioning: Beyond validity evidence based on internal structure. Psicothema. 2018;30(1):104-9.
35. Young JW. Ensuring valid content tests for English Language Learners. Educational Testing Service. 2008.
36. Wong S, Yang L, Riecke B, Cramer E, Neustaedter C. Assessing the usability of smartwatches for academic cheating during exams. En: Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2017. Association for Computing Machinery; 2017.
37. Bond L. Teaching to the Test: Coaching or Corruption. New Educ. 2008;4(3):216-23.
38. Lane S, Raymond M, Haladyna T. Handbook of Test Development [Internet]. 2nd ed. Lane S, Raymond M, Haladyna T, editores. International Journal of Testing. New York: Routledge; 2016. 676 p. Disponible en: http://www.tandfonline.com/doi/abs/10.1080/15305050701813433
39. Jurado A, Leenen I. Reflexiones sobre adivinar en preguntas de opción múltiple y cómo afecta el resultado del examen. Inv Ed Med. 2016;5(17):55-63.