El concepto moderno de validez y su uso en educación médica

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

Facultad de Medicina

Resumen

Para realizar inferencias apropiadas con base en los resultados obtenidos de las evaluaciones del aprendizaje en ciencias de la salud, es fundamental aportar evidencia de validez y así proveer el fundamento y la justificación de las decisiones que se tomen a partir de las evaluaciones. El concepto de validez es el más importante en evaluación educativa, pues aplica para todo tipo de uso de instrumentos de evaluación del aprendizaje, tanto sumativos como diagnósticos y formativos. En las últimas décadas han surgido nuevos marcos de referencia que modifican y enriquecen el concepto tradicional de validez. En este trabajo se exploran las perspectivas de Messick y Kane. Con respecto al primero se describen las fuentes de evidencia de validez y cómo obtenerlas, mientras que con relación al segundo se explican los pasos para llevar a cabo un argumento de usos que justifique las interpretaciones de los resultados de los exámenes. Con este panorama se presenta una perspectiva moderna de aproximación a la validez en evaluación educativa, de utilidad para los educadores en ciencias de la salud.

Palabras clave: *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: 16-octubre-2019. Aceptado: 2-diciembre-2019.
*Autora para correspondencia: Blanca Ariadna Carrillo Avalos. Av. Venustiano Carranza 2405, Col. Los Filtros, San Luis Potosí, S. L. P., México. CP 78210. Teléfono: 44 4826 2345, ext. 6635.
Correo electrónico: bariadna@gmail.com
La revisión por pares es responsabilidad de la Universidad Nacional Autónoma de México.
2007-5057/
https://doi.org/10.22201/facmed.20075057e.2020.33.19216

Current concepts of validity and its use in medical education

Abstract

In order to articulate appropriate inferences based on the scores obtained from learning assessments in the health sciences, the collection of validity evidence to support decisions made on the basis of these assessments is of central importance. The concept of validity is key in educational assessment, since it is used in all kinds of learning evaluation strategies: summative, diagnostic, and formative. In the last decades, new frameworks which modify and enhance the traditional concept of validity have emerged. In this paper, we explore the perspectives of Messick and Kane. Regarding the first one, we describe the sources of validity evidence and how to obtain them; and in regard to Kane's arguments, we explain the steps needed to state an argument of use that justifies the interpretations of the scores obtained from the assessments. This overview describes the current perspective to approach validity in educative assessment, useful for health sciences educators.

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

INTRODUCCIÓN

En una ocasión aplicamos un examen final de una materia de ciencias básicas que solo aprobó un pequeño porcentaje de alumnos; algunas personas comentaron que el examen no era válido, por lo que debíamos repetirlo. ¿Cómo podríamos comprobarlo? Primero, decir que un examen es válido o no, es un error de concepto frecuente que es importante despejar para contar con elementos que permitan elaborar y aplicar los exámenes de alto y bajo impacto, así como contar con resultados útiles.[1] Por otro lado, son numerosas las publicaciones que hablan acerca de aspectos de validez en evaluación en educación médica (como validez predictiva o validez de las preguntas del examen), cuyo análisis no menciona explícitamente el concepto actual de validez, y cómo se debe evaluar e interpretar.[2-4]

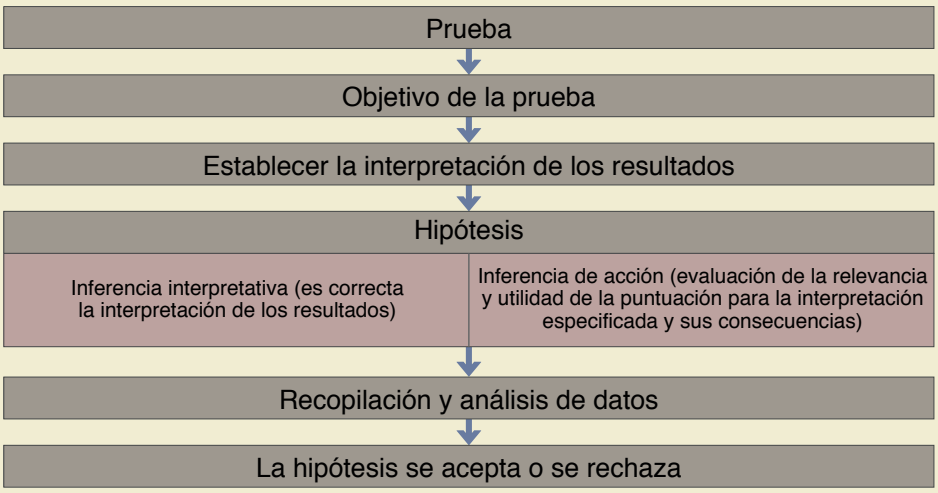
Al desarrollar y evaluar los exámenes, la validez, como el grado con que la evidencia empírica y las razones teóricas apoyan o refutan lo apropiado o adecuado de la interpretación o el uso que se da a los resultados de una evaluación, es la consideración más importante que debe hacerse.[5,6] Por otro lado, la característica o concepto que se mide en una evaluación específica es un constructo latente, y debe especificarse cuál es la interpretación que se va a dar acerca de éste con base en las puntuaciones obtenidas en la prueba. De esta manera son las inferencias que se hacen acerca de un constructo con base en la puntuación de una evaluación las que requieren evidencia de validez, mas no la evaluación por sí misma. Además, el análisis de la validez será en cuestión de grado y no de un enfoque dicotómico que certifique su existencia o inexistencia.[6-8]

El concepto de validez ha evolucionado desde la primera mitad del siglo XX, cuando se consideraba que un examen era válido cuando existía correlación con lo que pretendía medir.[9, citado por 10] Posteriormente se elaboró la teoría tradicional que identificaba tres tipos de validez: de contenido, de constructo y de criterio, esta última dividiéndose en validez concurrente y validez predictiva.[11] En las últimas décadas han surgido nuevos marcos de referencia que modifican y enriquecen el concepto tradicional de validez, de forma que actualmente existen dos de ellos que son considerados los más prominentes, por lo cual se estima necesario tomarlos en cuenta para evaluar la validez en evaluación educativa: el de Samuel Messick y el de Michael Kane. En este artículo presentamos una introducción general a estos marcos para la validez y sugerimos algunas ideas para su integración en educación en ciencias de la salud.

MARCO DE REFERENCIA DE MESSICK

El marco de referencia de Messick[8] considera que la validez de constructo es el único tipo que existe ya que las evaluaciones tienen como objetivo medir

Figura 1. Resumen del marco de referencia de Messick



Fuente: Elaboración propia.

constructos, es decir, las características o atributos de las personas que no pueden ser observados directamente (son latentes) y que se miden a través del examen diseñado.[6,11] Por ejemplo, el desempeño académico de un estudiante de medicina es una característica latente, por lo que se infiere a través de sus respuestas en los exámenes de cada asignatura, conformando un constructo susceptible de estudio.[7,12] A la luz de lo anterior, cualquier estudio de validez en el marco de Messick busca aportar, de forma directa o indirecta, evidencia para el constructo que subyace la evaluación.

Messick menciona que un análisis de la validez siempre parte de una hipótesis o inferencia acerca de la interpretación o el uso que se pretende dar a los resultados de la prueba. Posteriormente se deben recopilar y analizar los datos, enlazarlos a un marco teórico específico, y luego determinar la validez o invalidez de la hipótesis declarada para un momento particular en el tiempo, para una población específica **(figura 1)**.[7,8]

Así, este marco de referencia se enfoca en cinco fuentes de evidencia de validez. No es indispensable buscar todas estas fuentes en todos los análisis de resultados de exámenes. Las fuentes de evidencia de validez que se requieren dependen de los objetivos de la prueba y de sus consecuencias, entre otros aspectos[13], ya que éstas sirven para sustentar la interpretación que se haya determinado para la prueba previamente.[6,7] Por ejemplo, en el caso de pruebas de altas consecuencias como el examen de admisión a la escuela de medicina o el examen de titulación de enfermería, podría necesitarse mayor evidencia de validez que para una prueba utilizada con fines formativos.[6] A continuación, se discuten las cinco fuentes de evidencia de validez en el marco de Messick y algunos ejemplos de cómo documentarlas.

FUENTES DE EVIDENCIA DE VALIDEZ

1. Evidencia basada en el contenido de la prueba

El contenido de la prueba se refiere a los temas que evalúa; por ejemplo, en el caso de un examen de admisión abarcaría toda la información cuyo dominio debe demostrar un alumno antes de ingresar al nivel educativo que pretende. Este contenido también depende de las inferencias que se vayan a hacer a partir de las puntuaciones obtenidas en la prueba.[6]

Esta evidencia se puede obtener "a partir del análisis de la relación entre el contenido de la prueba y el constructo que pretende medir"[6], por ejemplo, se analiza la representatividad de la tabla de especifi-

caciones con respecto al dominio del conocimiento que se examina, las especificaciones del examen, representatividad de los ítems con respecto al dominio del conocimiento examinado, coincidencia del contenido de los ítems con las especificaciones del examen y relación lógica o empírica del contenido evaluado con el dominio del conocimiento que se examina.

Para documentar esta fuente de evidencia también se evalúan procesos de alineación, que evalúan la correspondencia entre el contenido de la prueba y los resultados de aprendizaje del alumno, es decir, qué tanto se representa el dominio del conocimiento en la prueba con base en criterios como la complejidad cognitiva, el currículo y los métodos instruccionales. Esto se puede lograr de diferentes formas, una de ellas consiste en que expertos califiquen la semejanza entre pares de ítems en términos de las habilidades y el conocimiento evaluados por medio de escalas tipo Likert.[6,14]

2. Evidencia basada en los procesos de respuesta

En los *Standards for Educational and Psychological Testing*[6] esta fuente de evidencia se refiere a que se puede comprobar la relación entre el constructo que se pretende medir y los procesos cognitivos que intervienen en la resolución de la tarea o los ítems de la prueba. Esta evidencia puede obtenerse por medio de entrevistas cognitivas, herramientas que permiten conocer la comprensión de términos clave, así como entender el razonamiento utilizado para llegar a la respuesta correcta y así evitar falsos positivos (llegar a la respuesta correcta después de un razonamiento erróneo), de manera que el sustentante realmente esté aplicando lo necesario para resolver el problema propuesto y que así logre obtener resultados favorables en otros contextos.[15] También existen modelos matemáticos que relacionan la dificultad de los ítems o el tiempo de respuesta con los procesos cognitivos hipotéticos, mismos que permiten aportar evidencia de este tipo.[16]

Cabe mencionar que Downing[7] incluye para esta fuente de evidencia de validez también un análisis de aspectos asociados con la administración del examen, por ejemplo, la familiaridad de los sustentantes con el formato del examen, que sepan llenar adecuadamente las hojas de respuesta, la claridad de las instrucciones, etc. Sin embargo, es importante aclarar que esta interpretación de Downing[7] se encuentra algo desalineada con la visión del mismo Messick y de los psicómetras prominentes en esta área, como Kane y Embretson, entre otros.

3. Evidencia basada en la estructura interna

La estructura interna es el grado en que las relaciones de los ítems de la prueba están alineadas con la teoría detrás del constructo que se mide.[6] Evidencia de este tipo se puede obtener analizando las características psicométricas de las preguntas del examen, las características de la escala, y el modelo psicométrico que se utilizó para establecer la escala y calificar el examen.[7] El análisis de datos para obtener evidencia de validez de este tipo suele recurrir a análisis factorial (exploratorio o confirmatorio) o análisis en el marco de la teoría de respuesta al ítem; ambos permiten investigar las relaciones entre las respuestas en los ítems y el constructo subyacente a la prueba.[17,18]

El análisis de la estructura interna también atañe a la confiabilidad; en general, es importante documentar que las puntuaciones pudieran ser reproducibles si se aplicara nuevamente la prueba. De lo contrario, la interpretación de los resultados de este examen se puede ver comprometida.[7,18,19]

4. Evidencia basada en las relaciones con otras variables

Este tipo de evidencia se basa en el análisis de la relación de los resultados de la prueba con los resultados de otras pruebas que midan o no el mismo constructo u otras variables externas a la prueba. Proporciona información acerca del grado en que estas relaciones son coherentes con el constructo en el que se basan las interpretaciones de los resultados de la prueba.[6] Se puede buscar evidencia por esta fuente con base en relaciones convergentes (cuando se evalúan las relaciones entre las puntuaciones y medidas del mismo constructo) y/o discriminantes (cuando se evalúan las relaciones entre las puntuaciones y medidas de constructos diferentes).[7] Una manera de investigar ambos tipos de relaciones es a través de una matriz multirasgo-multimétodo, que es una matriz de correlaciones entre distintas prue-

bas que, en conjunto, miden dos o más constructos a través de dos o más métodos.[20]

Se consideran dos diseños para la evidencia de validez de este tipo:[6]

- Estudio predictivo. Evalúa el grado de la relación entre las puntuaciones de la prueba y las puntuaciones del criterio que se obtiene en un tiempo posterior. Por ejemplo, estudios que evalúan exámenes de admisión académica y que investigan la relación con el desempeño académico subsecuente.
- Estudio concurrente. Evalúa el grado de la relación entre las puntuaciones de la prueba y las puntuaciones del criterio que se obtiene al mismo tiempo. En este tipo de estudios se evitan los cambios temporales y pueden ser útiles para buscar formas alternas de medición del constructo en cuestión, por ejemplo, analizar la correlación de los puntajes de una variante corta de una prueba con los de una variante original más larga, que mide el mismo constructo, pero ya cuenta con evidencia de validez.

La generalización de los resultados que aporta el estudio de esta fuente de validez depende de que las condiciones en la nueva situación sean iguales a las presentes en el análisis original. Los resúmenes estadísticos de los estudios de validación anteriores en condiciones semejantes, como en un meta-análisis, pueden ser útiles para estimar las nuevas relaciones, pero dependen del tamaño de la muestra y de la cantidad de estudios realizados a lo largo del tiempo.[6,21]

5. Evidencia basada en las consecuencias de la prueba

Generalmente, la interpretación y el uso de los resultados de la prueba tienen impacto o consecuencia de diferentes grados o tipos sobre los sustentantes. Por ejemplo, en el caso de las evaluaciones de admisión para una licenciatura, esta evidencia lleva a reflexionar sobre las posibles equivocaciones en la interpretación de los resultados de la prueba con respecto a falsos positivos y falsos negativos, así como tomar en cuenta estas consecuencias negativas para que se lleve a cabo una evaluación de qué tan grave es un falso positivo y qué tan grave un falso negativo y que se considere al ponderar las consecuencias diferenciales de ambos tipos de errores.

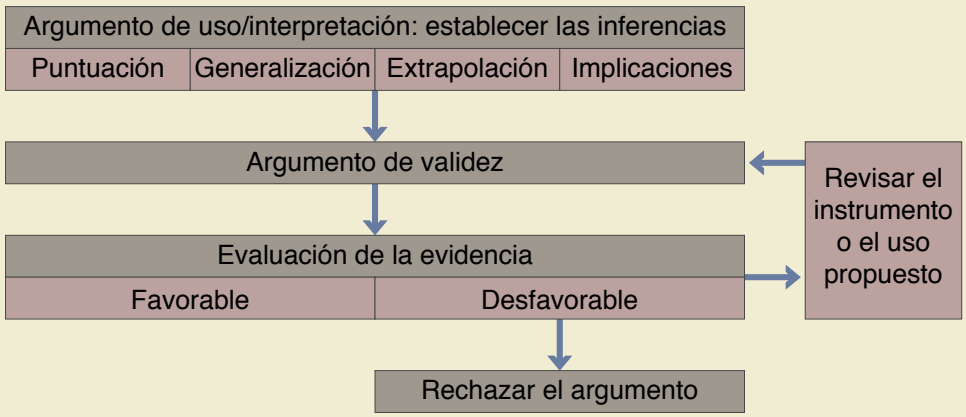
Esta fuente de validez requiere analizar el impacto de los resultados de la prueba en los estudiantes y la sociedad, el balance entre las consecuencias positivas y las negativas involuntarias, lo razonable del punto de corte de aprobado/reprobado o admitido/no admitido, las consecuencias de aprobar o reprobar, de los falsos positivos y falsos negativos, y las consecuencias institucionales y del estudiante.[6,7] Este análisis puede realizarse por medio de entrevistas y grupos focales, así como la teoría de acción para identificar los componentes críticos de los programas académicos y sus puntos de impacto.[22]

Como ejemplo, considérese el Examen Nacional para Aspirantes a Residencias Médicas, que "es un instrumento de medición de conocimientos en el contexto del ejercicio de la medicina general, objetivo y consensuado, que constituye la primera etapa del proceso para ingresar al Sistema Nacional de Residencias Médicas."[23] A pesar del objetivo establecido por los desarrolladores de esta evaluación, algunas instituciones utilizan sus resultados como una forma de determinar cual es "la mejor escuela de medicina" en nuestro país, produciendo consecuencias no intencionadas e indeseables. Analizar estas consecuencias y hacer lo necesario para evitarlas en la medida de lo posible constituye un ejemplo de este tipo de evidencia de validez.

MARCO DE REFERENCIA DE KANE

Kane consideró que, aunque la visión de Messick acerca de la validez de constructo es importante, no es fácil de evaluar, ya que no provee de guías para iniciar el procedimiento, y no es muy práctica[24]; por ello desarrolló su propio marco de referencia que se enfoca en el proceso de recolección de evidencia de validez mediante cuatro inferencias para desarrollar un argumento de validez.[25] El planear un examen considerando las fuentes de validez marca el camino para partir de la evaluación de una sola observación (inferencia de puntuación) hacia la puntuación general del examen (generalización) y de ahí a establecer las implicaciones de la puntuación en el desempeño en la vida real (extrapolación), llegando finalmente a la interpretación de esta información y a la toma de decisiones (implicaciones).[26] Una ventaja de este

Figura 2. Marco de referencia de Kane



Basado en Cook et al., 2015.

acercamiento a la validez es que es factible para quienes no poseen experiencia amplia en psicometría, además de que propone pasos muy claros.[27]

En general, los pasos que propone son dos: el primero es establecer el argumento de uso o interpretación (AUI) y el segundo es desarrollar el argumento de validez; este último es facilitado al considerar los cuatro tipos de inferencias **(figura 2)**.

1. Establecer el argumento de uso o interpretación (AUI)

La interpretación de los resultados de la prueba implica explicar el significado de la puntuación, mientras que el uso de las puntuaciones se refiere a las decisiones que se toman con base en los resultados de la prueba. Kane considera que ambos términos (interpretación y usos) incluyen todas las suposiciones que se pueden hacer al respecto de las puntuaciones de una prueba, por lo que se debe establecer la validez de la interpretación o el uso de las puntuaciones en términos de lo creíble y apropiado que tengan en un punto del tiempo. Tener claro lo que se quiere evaluar permite elaborar un plan de evaluación preciso, por lo que el AUI puede conformar una red de inferencias y suposiciones que van desde el desempeño en las pruebas hasta las conclusiones que se obtienen, y las decisiones que se toman con base en estas conclusiones.[28,29]

Kane sugiere las siguientes inferencias que se encuentran presentes en la mayoría de los AUI, aunque también menciona que no es indispensable evaluarlas todas:[29,30]

- Inferencia de puntuación. Es la suposición acerca de lo apropiado de los criterios de la puntuación y las reglas para combinar las puntuaciones, mismas que conforman un estimado acerca de un atributo y son la base para la toma de decisiones.
- Inferencia de generalización. Si la prueba contiene una muestra de posibles escenarios o posibles ítems, esta inferencia supone que el sustentante va a obtener puntuaciones semejantes al presentar otra prueba con ítems diferentes extraídos del mismo universo de ítems, de manera que las puntuaciones observadas son representativas de todo el universo de puntuaciones posibles. Esta inferencia puede utilizar evidencia empírica en el marco de la teoría de la generalizabilidad,[31] debido a la importancia de puntuaciones reproducibles y generalizables.
- Inferencia de extrapolación. Por medio de este tipo de suposiciones se podría extender la interpretación a otros dominios de desempeño y predecir cuál será el resultado del sustentante en contextos diferentes al del examen. Un ejemplo de este tipo de inferencia sería que si la puntua-

ción observada tiene un valor particular (examen de admisión), entonces se espera un valor específico del criterio (desempeño académico durante la carrera); las herramientas analíticas para evaluar inferencias de este tipo suelen utilizar modelos de regresión.

- Inferencia de implicaciones. Se refiere al impacto que tiene la interpretación de los resultados de la prueba en el sustentante, en su familia y en la sociedad. Kane considera que, si las consecuencias de la interpretación de los resultados de una prueba son negativas, entonces la prueba no debería utilizarse.

2. Establecer el argumento de validez

Una vez que se han establecido las inferencias concernientes a las puntuaciones de la prueba en cuestión, se deben evaluar las garantías o métodos de comprobación de estas inferencias. Por ejemplo, la garantía de una inferencia de extrapolación con interés predictivo sería una ecuación de regresión, cuyo soporte estaría conformado por un análisis empírico acerca de la relación entre la puntuación de la prueba y los resultados del criterio seleccionado. El calificador de la garantía es el término que expresa la fuerza de la relación que se está analizando, y puede expresarse de manera numérica y con palabras (como coeficientes de correlación).[29]

Con estas consideraciones, el primer paso será realizar un análisis conceptual del AUI y verificar que sea coherente y que todas las inferencias importantes se encuentren presentes. Posteriormente, se deberán evaluar las inferencias presentadas. En la **tabla 1** se resumen las inferencias que propone Kane, así como los procedimientos que se deben definir y la manera de evaluarlos.

El argumento de validez debe ser claro para poder ser reproducible por cualquier investigador, conteniendo detalles específicos y presentando información coherente, de manera que las conclusiones sean lógicas. Por lo anterior, el argumento también debe estar completo y ser verificable.[32]

CONCLUSIONES

Se han revisado brevemente los marcos de referencia modernos y prominentes de validez a considerar cuando se interpretan y utilizan los resultados de las pruebas evaluativas en medicina; esta información es importante ya que su conocimiento y aplicación permitirá iniciar la elaboración de evaluaciones mejor planeadas y con objetivos más claros, además de que los resultados serán realmente útiles y su interpretación tendrá mayor grado de validez. No todas las fuentes de evidencia de validez se encontrarán presentes en todos los exámenes; sin embargo, son indispensables las que sustenten la interpretación descrita al inicio de la planeación.

Por otro lado, mientras que el marco de referencia de Messick deja claras las fuentes de evidencia de validez, Kane propone los pasos para que, a partir de inferencias bien definidas, podamos analizar estas fuentes. Al realizar cualquier análisis de validez es importante hacer referencia al marco que se está utilizando y explicar la justificación de las fuentes de evidencia propuestas, las que deben estar alineadas al uso e interpretaciones establecidos. Ambos marcos de referencia toman en cuenta aspectos semejantes de las evaluaciones, por lo que una posible línea de investigación sería considerar las fuentes de evidencia de validez de Messick como pruebas o garantías de las inferencias que se hacen a partir del método de Kane, obteniendo así las fuentes de evidencia de validez de manera sistematizada.

REFERENCIAS

1. Sánchez-Mendiola M. «Mi instrumento es más válido que el tuyo»: ¿Por qué seguimos usando ideas obsoletas? Inv Ed Med. 2016;5(19):133-5.
2. Roméu Escobar MR, Díaz Quiñones JA. Valoración metodológica de la confección de temarios de exámenes finales de Medicina y Estomatología. Rev Cuba Educ Med Super. 2015;29(3):522-31.
3. Salvatori P. Reliability and Validity of Admissions Tools Used to Select Students for the Health Professions. Adv Heal Sci Educ. 2001;6(2):159-75.
4. Baladrón J, Curbelo J, Sánchez-Lasheras F, Romeo-Ladrero JM, Villacampa T, Fernández-Somoano A. El examen al examen MIR 2015. Aproximación a la validez estructural a través de la teoría clásica de los tests. FEM. 2016;19(4):217.
5. Shepard LA. Evaluating test validity: reprise and progress. Assess Educ. 2016;23(2):268-80. Disponible en: http://dx.doi.org/10.1080/0969594X.2016.1141168
6. 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

Tabla 1. Las inferencias y sus fuentes de evidencia correspondientes para establecer el argumento de validez

Inferencia	Consiste en	Procedimientos a definir, establecer o seleccionar	Evaluación empírica de:
Puntuación	Suposición acerca de lo apropiado de los criterios de la puntuación y las reglas para combinar las puntuaciones.	• Ítems y opciones de respuesta (preguntas de opción múltiple, falso/verdadero) • Formato de la observación • Estandarización entre formatos y ocasiones • Rúbrica o criterio de puntuación, procedimientos de implementación, estándar de aprobado/no aprobado • Selección y entrenamiento de los evaluadores (p ej., ECOE) • Reglas para combinar los elementos relacionados con la prueba a partir de fuentes diferentes o para separar elementos no relacionados de la misma fuente • Seguridad de los datos y control de calidad	• Desempeño de ítems y de opciones de respuesta • Formato de observación • Estandarización • Rúbrica o criterio de puntuación • Selección y entrenamiento de los evaluadores, confiabilidad y precisión de los evaluadores (p ej. en evaluación de desempeño – ECOE) • Seguridad de los datos y control de calidad
Generalización	Los ítems de la prueba conforman una muestra del universo de ítems posibles. Esta inferencia supone que se puede generalizar hacia todo el universo de ítems posibles. Se relaciona con la confiabilidad.	• Estrategia de muestreo de los ítems • Tamaño de la muestra (número de preguntas)	• Confiabilidad o generalizabilidad por medio de la teoría de la generalizabilidad • Teoría de respuesta del ítem
Extrapolación	Se podría extender la interpretación a otros dominios de desempeño y predecir cuál será el resultado del sustentante en contextos diferentes al del examen o tareas diferentes en contextos diferentes.	• Alcance de la prueba • Autenticidad del contexto de la prueba • Autenticidad del ítem/escenario • Análisis que demuestren la relación entre el desempeño en la prueba y los dominios o contextos diferentes a los que se desea extrapolar	• Análisis para definir el alcance/objetivos • Acuerdo entre el proceso y el constructo • Relevancia y autenticidad • Correlación con otra medida que presente la misma relación esperada (con referencia al criterio o convergente; concurrente o predictiva) • Discriminación • Sensibilidad al cambio después de la intervención • Perfil del constructo • Funcionamiento diferencial del ítem
Implicación	Acerca del impacto de la interpretación de los resultados de la prueba sobre el sustentante, otros interesados y la sociedad.	• Estándar de aprobado/no aprobado • Acciones planeadas con base en los resultados de la prueba • Consecuencias voluntarias o involuntarias de las decisiones que se toman a partir de los resultados de la prueba	• Estándar de aprobado/no aprobado • Efectividad de las acciones basadas en los resultados de la prueba • Consecuencias voluntarias o involuntarias de la prueba • Funcionamiento diferencial del ítem

Fuente: Cook et al., 2015; Kane, 2013; Schuwirth & van der Vleuten, 2012.

Association, American Psychological Association & National Council on Measurement in Education; 2014. 243 p.
7. Downing SM. Validity: On the meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-7.
8. Messick S. Validity. 1987. Disponible en: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2330-8516.1987.tb00244.x
9. Guilford JP. New Standards For Test Evaluation. Educ Psychol Meas. 1946;6(4):427-39.
10. Shepard LA. Evaluating Test Validity." En: Darling-Hammon L, editor. Review of Research in Education. Washington, DC.: AERA; 1993. p. 405-50.
11. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955;52(4):281-302.
12. York TT, Gibson C, Rankin S. Defining and measuring academic success. PARE. 2015;20(5):1-20.
13. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. Adv Simul. 2016;1(1):1-12. Disponible en: http://dx.doi.org/10.1186/s41077-016-0033-y
14. Sireci S, Faulkner-Bond M. Evidencia de validez basada en el contenido del test. Psicothema. 2014;26(1):100-7.
15. Padilla JL, Benítez I. Evidencia de validez basada en los procesos de respuesta. Psicothema. 2014;26(1):136-44.
16. Embretson SE. A Cognitive Design System Approach to Generating Valid Tests: Application to Abstract Reasoning. Psychol Methods. 1998;3(3):380-96.
17. 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.
18. Rios J, Wells C. Evidencia de validez basada en la estructura interna. Psicothema. 2014;26(1):108-16.
19. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. 2004;38:1006-12.
20. Campbell D, Fiske D. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull. 1959;56(2):81-105.
21. Coates H. Establishing the criterion validity of the Graduate Medical School Admissions Test (GAMSAT). Med Educ. 2008;42(10):999-1006.
22. Lane S. Evidencia de validez basada en las consecuencias del uso del test. Psicothema. 2014;26(1):127-35.
23. Secretaría de Salud, Secretaría de Educación Pública, Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud. XLIII Examen Nacional para Aspirantes a Residencias Médicas. Convocatoria 2019. Ciudad de México, México.; 2019. Disponible en: http://www.cifrhs.salud.gob.mx/site1/enarm/docs/2019/E43_convo_2019.pdf
24. Kane M. Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. Lang Test. 2011;29(1):3-17.
25. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: Validity evidence for qualitative educational assessments. Acad Med. 2016;91(10):1359-69.
26. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: A practical guide to Kane's framework. Med Educ. 2015;49(6):560-75.
27. Brennan R. Commentary on "Validating the Interpretations and Uses of Test Scores." J Educ Meas. 2013;50(1):74-83.
28. Kane MT. An argument-based approach to validity in evaluation. Psychol Bull. 1992;112(3):527-35.
29. Kane MT. Validating the Interpretations and Uses of Test Scores. J Educ Meas. 2013;50(1):1-73.
30. Chalhoub-Deville M. Validity theory: Reform policies, accountability testing, and consequences. Lang Test. 2016;33(4):453-72.
31. Brennan R. Generalizability Theory. New York: Springer-Verlag New York; 2001. XX, 538.
32. Schuwirth LWT, van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. Med Educ. 2012;46(1):38-48.