XVI Congreso Nacional de Investigación Educativa
CNIE-2021

Un acercamiento a la unificación de los marcos de referencia de validez de Messick y de Kane

Blanca Ariadna Carrillo Avalos
Facultad de Medicina, Universidad Autónoma de San Luis Potosí
bariadna@gmail.com

Iwin Leenen
Facultad de Psicología, Universidad Nacional Autónoma de México
iwin.leenen@gmail.com

Melchor Sánchez Mendiola
Coordinación de Universidad Abierta, Innovación Educativa y Educación a Distancia, Universidad Nacional Autónoma de México
melchor_sanchez@cuaieed.unam.mx

Área temática 12. Evaluación educativa.
Línea temática: Aportes metodológicos a la evaluación educativa.
Tipo de ponencia: Aportaciones teóricas.

Resumen

La validez es el grado en que la evidencia apoya o refuta si la interpretación que se le da a las puntuaciones de las evaluaciones es adecuada. Dos marcos de referencia contemporáneos de validez son el de Samuel Messick y el de Michael Kane. El trabajo del primero es el más aceptado, se encuentra descrito en los *Estándares para pruebas educativas y psicológicas*, y contempla cinco fuentes de evidencia de validez; sin embargo, no especifica el orden para obtener esta información o cómo debe interpretarse, lo que implica un análisis complejo y poco práctico. El de Kane se enfoca en dos pasos que permiten obtener evidencia de validez para probar cuatro supuestos; aunque es más ordenado, tampoco queda muy claro cómo se debe interpretar esta información para establecer el grado de validez. El objetivo de este trabajo fue unificar ambos marcos de referencia en una propuesta de método para probar la validez de la interpretación de las puntuaciones. Se compararon ambos marcos de referencia en cuanto a la teoría, los procedimientos y las fuentes de evidencia propuestos; posteriormente, se unieron las fuentes de evidencia de validez de Messick con los supuestos de Kane para las que podían aportar garantías. Finalmente, se establecieron los pasos a desarrollar con base en el marco de Kane. El resultado es un método sistemático para validar la interpretación de los resultados de las evaluaciones, el que resulta de importancia e interés principalmente para pruebas de altas consecuencias como exámenes de admisión o de titulación.

Palabras clave: *Validez, evaluación del aprendizaje, pruebas de altas consecuencias.*

Introducción

El marco de referencia tradicional consideraba tres tipos de validez: de contenido, de constructo y de criterio; la validez de criterio se dividía a su vez en concurrente y en predictiva. (Cronbach & Meehl, 1955). Actualmente existen dos marcos de referencia contemporáneos que consideran a la validez como el grado en que la evidencia apoya o refuta si la interpretación que se le da a las puntuaciones de las evaluaciones es adecuada (Downing, 2003): el de Samuel Messick y el de Michael Kane. El trabajo del primero es el más aceptado y se encuentra descrito en los Estándares para pruebas educativas y psicológicas, y contempla cinco fuentes de evidencia de validez (American Educational Research Association et al., 2018; Messick, 1989, 1990); sin embargo, no especifica el orden para obtener esta información o cómo debe interpretarse, lo que implica un análisis complejo y con frecuencia poco práctico. El de Kane se enfoca en dos pasos que permiten obtener evidencia de validez para probar cuatro supuestos; aunque es más ordenado, tampoco queda claro cómo se debe interpretar esta información para establecer el grado de validez. (Kane, 1992, 2016). El objetivo de este trabajo fue unificar ambos marcos de referencia en una propuesta de método, para probar la validez de la interpretación de las puntuaciones en evaluaciones educativas, considerando las fuentes de evidencia de validez de Messick como garantías para probar los supuestos de Kane.

Desarrollo

Se llevó a cabo un estudio exploratorio que comprendió cuatro fases: comparación de ambos marcos de referencia en cuanto a la teoría; comparación de los procedimientos; comparación de las fuentes de evidencia propuestos; en la última fase se unieron las fuentes de evidencia de validez de Messick con los supuestos de Kane para las que podían aportar garantías, integrándose en los pasos generales de Kane.

1. Marcos teóricos

- Messick. - considera que las evaluaciones tienen como objetivo medir constructos (características latentes que no pueden ser observadas directamente y que se miden por medio de un examen), por lo que propone que la validez es un concepto unitario y holístico, que corresponde a la de constructo, y que esta se alimenta de diferentes fuentes. (American Educational Research Association et al., 2018; Cronbach & Meehl, 1955).
- Kane. – Brennan (2013) observó que el problema principal de aproximación mediante fuentes de evidencia de validez es que no están probando hipótesis específicas. Kane propone el desarrollo de un argumento de usos e interpretaciones de las puntuaciones con base en cuatro supuestos: puntuación, generalización, extrapolación e implicaciones. (Kane, 1992). También menciona que la validez de la interpretación o del uso de las puntuaciones de las pruebas se puede definir en términos de lo apropiado del uso o la interpretación en un punto específico en el tiempo. (Kane, 2013).

2. Procedimientos de análisis de validez

- Messick. - indica que, para iniciar un análisis de esta naturaleza, se debe establecer una hipótesis acerca del uso que se pretende dar a la prueba en cuanto a un tiempo y población específicos; luego, reunir y analizar los datos que demuestren la hipótesis, enlazarlos al marco teórico y determinar si la hipótesis se acepta o se rechaza. Con respecto a los datos que servirán como garantía para la hipótesis planteada, el autor describió cinco fuentes de evidencia de validez; su presencia depende de los objetivos de la prueba y de sus consecuencias. Estas fuentes son evidencias basadas en el contenido de la prueba, los procesos de respuesta, la estructura interna, las relaciones con otras variables y las consecuencias de la prueba. (American Educational Research Association et al., 2018; Messick, 1989, 1990).
- Kane. - se enfoca en dos pasos:
  I. Establecer el argumento de usos e interpretaciones de la prueba. La interpretación se refiere a la explicación del significado de la puntuación de la prueba, mientras que el uso considera las decisiones que se toman con base en la puntuación de la prueba. Estos dos conceptos son analizados a través de cuatro supuestos: puntuación, generalización, extrapolación e implicaciones.
  II. Establecer el argumento de validez al analizar los datos que pueden probar estos cuatro supuestos y determinar si la evidencia es favorable o desfavorable hacia un grado de validez aceptable. (Kane, 1992, 2013).

3. Fuentes de evidencia

- Messick. - Las cinco fuentes de evidencia de validez y sus garantías correspondientes se resumen en la tabla 1.
- Kane. - Los supuestos y sus fuentes de evidencia correspondientes para establecer el argumento de validez se describen en la tabla 2. (Cook et al., 2015; Kane, 2013; Schuwirth & van der Vleuten, 2012).

Puntos en común entre los marcos de referencia de Messick y de Kane:

En la figura 1 se representan los pasos de evaluación de la validez de la interpretación de la prueba que propone Kane, junto con los supuestos que deben verificarse. Debajo de cada tipo de supuesto se ha anotado la fuente de validez de constructo según Messick, considerando que, de encontrarse, constituirán la garantía para los supuestos de Kane. Al final se obtendría la determinación de aceptación o rechazo de los supuestos establecidos al principio, es decir, la aceptación o rechazo de la validez de la interpretación de la prueba.

4. Propuesta de método

Con base en los puntos en común identificados entre ambos marcos de referencia y en el trabajo publicado por Tavares et al, (2018) hemos desarrollado la propuesta de método (figura 2) que se describe a continuación:

1.a. Formular el argumento de usos e interpretaciones

Acerca del objetivo o la interpretación de los resultados de la prueba. En este paso se describen los puntos a continuación:

a) El objetivo de la prueba

b) Los usuarios propuestos

c) Los usos

d) El constructo medido

e) Las interpretaciones de los resultados: con referencia a norma o a criterio.

f) La población examinada

- Frecuencias.
  - Hombres/mujeres
  - Edad
- Medidas de tendencia central.
- Medidas de dispersión.

1.b. Elaborar las hipótesis a probar según el tipo de supuesto

Las hipótesis que pueden funcionar como guía para elaborar las hipótesis propias, según los supuestos, han sido publicadas por varios autores. Se anotan en la tabla 3.

2.a. Crear un plan para probar las hipótesis

Se establece un plan para identificar y analizar las fuentes de validez necesarias para probar cada una de las hipótesis determinadas en el paso anterior (tabla 4).

2. b. Evaluar la evidencia y formular un juicio

Describir, por medio del desarrollo de un argumento de validez, si las hipótesis se aceptan o se rechazan para el uso propuesto. Por medio de esta discusión se determina el grado de validez del examen analizado.

Conclusiones

La validez es uno de los pilares del análisis de las pruebas de evaluación, además de la confiabilidad y la justicia. Entre los obstáculos para su evaluación se encuentran que no existe todavía consenso generalizado

sobre su definición o sobre los procedimientos para evaluarla; sin embargo, los marcos de referencia de Messick y de Kane proveen de información útil y práctica para llevar a cabo el proceso de validación.

El resultado de este trabajo es la propuesta de un método sistemático para validar la interpretación de los resultados de las evaluaciones, que integra ambos modelos conceptuales. La validación es un proceso que habitualmente requerirá trabajo de equipo para realizarlo de manera completa y expedita, y cuyos resultados cobran mayor relevancia en el caso de las pruebas de altas consecuencias, como exámenes de admisión o de titulación, debido al escrutinio al que están sometidas por las implicaciones que tienen sobre todos los usuarios: sustentantes y sus familias, instituciones, académicos, entre otros.

Tablas y figuras

Tabla 1. Fuentes de evidencia de validez, definición y posibles fuentes documentales

Fuente de evidencia	Definición:	Posibles fuentes documentales
1. Evidencia basada en el contenido de la prueba.	Se obtiene a partir del "… análisis de la relación entre el contenido de la prueba y el constructo que pretende medir."	Representatividad de la tabla de especificaciones con respecto al dominio del conocimiento que se examina, 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.
2. Evidencia basada en los procesos de respuesta.	Es la correspondencia adecuada entre el constructo que evalúa la prueba y los procesos que ejecutan los sustentantes al responder.	Entrevistas cognitivas, modelos matemáticos que relacionan la dificultad de los ítems o el tiempo de respuesta con los procesos cognitivos hipotéticos. Downing incluye la familiaridad de los sustentantes con el formato del examen, que sepan llenar adecuadamente las hojas de respuesta, la claridad de las instrucciones.
3. Evidencia basada en 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.	Las características psicométricas de las preguntas del examen, las características de la escala, el modelo psicométrico que se utilizó para establecer la escala y calificar el examen, la confiabilidad, generalizabilidad, funcionamiento diferencial del ítem, etc.
4. Evidencia basada en las relaciones con otras variables.	Se obtiene a partir del análisis de la relación entre el constructo con variables externas, por ejemplo, otras pruebas que midan el mismo constructo.	Las 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).
5. Evidencia para la validez y las consecuencias de la prueba.	Se refiere a que la interpretación de los resultados de la prueba tiene impacto de diferentes grados de consecuencia sobre el sustentante.	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.

Fuente: Downing, 2003; Embretson, 1998, 1999; Messick, 1989, 1990.

Tabla 2. Los supuestos y sus fuentes de evidencia correspondientes para establecer el argumento de validez

Supuesto	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 • 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 • 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 • Seguridad de los datos y control de calidad
Generalización	Los ítems de la prueba conforman una muestra del universo de ítems posibles. Este supuesto supone que se puede generalizar hacia todo el universo de ítems posibles. Se relaciona con la confiabilidad.	• Estrategia de muestreo • Tamaño de la muestra	• Confiabilidad o 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 para cambiar 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 entera.	• 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 (p ej. curva ROC) • 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.

Tabla 3. Hipótesis generales que pueden desarrollarse en el AUI

Supuestos	Hipótesis asociadas a los componentes de Kane
Puntuación	Las preguntas fueron administradas bajo condiciones estandarizadas – la regla es apropiada. Las puntuaciones fueron registradas de manera rigurosa – la regla se aplica como se especificó. Los algoritmos de puntuación fueron aplicados correctamente – la puntuación está libre de sesgo. Se implementaron los procedimientos de seguridad apropiados.
Generalización	¿Cuáles son las fuentes de medición del error que contribuyen a las puntuaciones observadas en la evaluación? ¿Qué tan semejantes serían las puntuaciones entre las réplicas del procedimiento de medición? ¿En qué medida se utilizó un proceso sistemático para construir las formas de la prueba? La muestra es representativa del universo de observaciones posibles. La muestra es lo suficientemente grande como para controlar para el error aleatorio. La puntuación puede generalizarse de la muestra a la población específica: ítems, jueces, etc.
Extrapolación	La puntuación observada se relaciona con el constructo de la vida real de interés. No hay probables errores sistemáticos que socaven la extrapolación. Las puntuaciones predicen los resultados de la vida real de interés. Existen aspectos artificiales de las condiciones de la prueba que impacten en las puntuaciones.
Implicaciones	La puntuación de corte fue establecida de manera razonable. Las implicaciones (interpretaciones) son apropiadas. Las propiedades de las puntuaciones apoyan las implicaciones (interpretaciones) asociadas.

Fuente: Clauser, Margolis, & Swanson, 2008; Hatala, Cook, Brydges, & Hawkins, 2015.

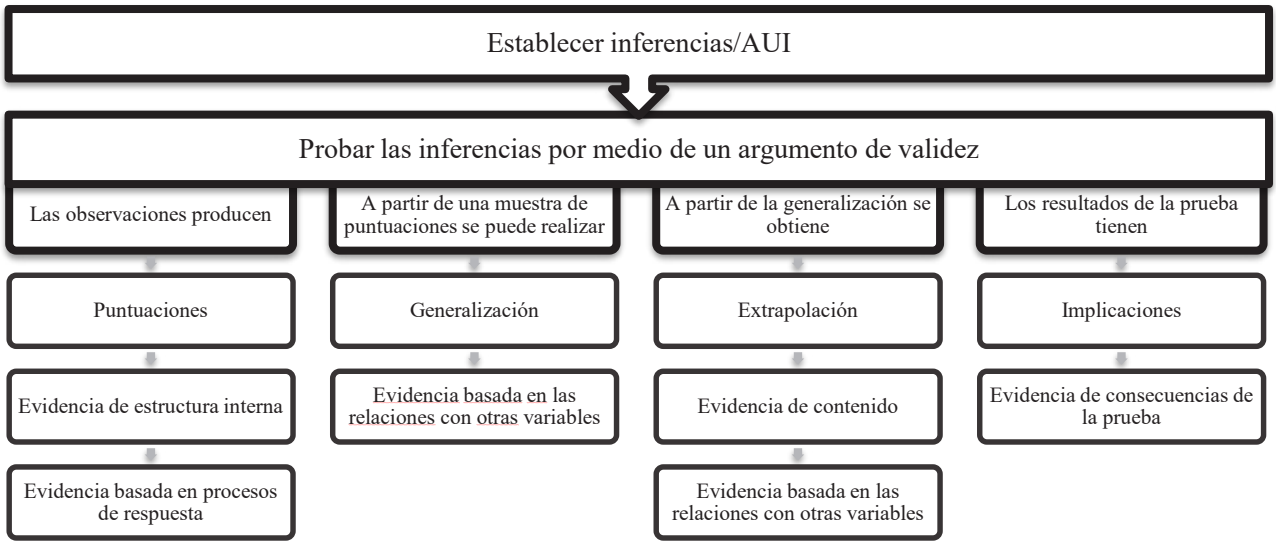
Tabla 4. Fuentes de evidencia de validez de Messick para comprobar las hipótesis con base en Kane.UAM-C

Supuesto (Kane)	Fuente de validez (Messick)	Procedimientos a definir, establecer o seleccionar	Evaluación empírica de:
Extrapolación	De contenido	1. Diseño del examen 2. Representatividad del diseño del examen con los dominios del conocimiento 3. Especificaciones del examen 4. Relación del contenido de los ítems con el diseño del examen 5. Representatividad de los ítems con el dominio del conocimiento 6. Relación lógica del contenido evaluado con el dominio del conocimiento 7. Calidad de las preguntas 8. Calificaciones de quienes elaboran los componentes del examen 9. Revisión de sensibilidad cultural 10. Alcance de la prueba 11. Autenticidad del contexto de la prueba 12. Autenticidad del ítem/escenario	1. Instructivo del examen: contenido, características, tiempo para contestar, número de preguntas en total y por tema evaluado, acceso al instructivo por parte de los sustentantes. 2. Relación de lo que se pregunta con los dominios del conocimiento establecidos para ser evaluados. 3. Características de quienes elaboran el examen: certificaciones y entrenamiento. 4. Análisis para definir el alcance/objetivos 5. Acuerdo entre el proceso y el constructo 6. Relevancia y autenticidad 7. Correlación con otra medida que presente la misma relación esperada (con referencia al criterio o convergente; concurrente o predictiva) 8. Discriminación 9. Sensibilidad al cambio después de la intervención 10. Definición del constructo evaluado 11. Funcionamiento diferencial del ítem

Puntuaciones	Procesos de respuesta	1. Familiaridad de los sustentantes con el formato 2. Control de calidad de la máquina de escaneo de calificaciones 3. Validación de la clave en exámenes piloto 4. Exactitud al combinar los diferentes formatos de puntuación 5. Control de calidad y exactitud de las puntuaciones finales 6. Análisis de subpuntuaciones 7. Exactitud de aplicación de las reglas de decisión de aprobado/no aprobado a las puntuaciones 8. Control de calidad del reporte de la puntuación a los estudiantes y profesores 9. Descripción o interpretación entendible y exacta de las puntuaciones para los estudiantes. 10. Ítems y opciones de respuesta (preguntas de opción múltiple, falso/verdadero) 11. Estandarización entre formatos y ocasiones 12. Rúbrica o criterio de puntuación, procedimientos de implementación, estándar de aprobado/no aprobado 13. Reglas para combinar los elementos relacionados con la prueba a partir de fuentes diferentes o para separar elementos no relacionados de la misma fuente 14. Seguridad de los datos y control de calidad	1. Instructivos de los componentes de la prueba: explicación de los procesos que evalúan y cómo lo hacen, cuál es el formato de aplicación de la prueba. 2. Explicación de la manera de calificar la prueba. 3. Evidencia de certificación de las máquinas de escaneo 4. Evidencia de los resultados de las pruebas piloto y los cambios realizados con base en estas 5. Justificación de ponderaciones al combinar los diferentes formatos de calificación 6. Formatos de reporte de calificaciones: que incluyan todas las puntuaciones y explicación de las ponderaciones de manera clara y en concordancia con el instructivo . 7.Desempeño de ítems y de opciones de respuesta 8.Estandarización 9.Rúbrica o criterio de puntuación 10. Seguridad de los datos y control de calidad
Puntuaciones Generalización	Estructura interna	1. Datos de análisis del ítem a. Índice de dificultad o discriminación del ítem b. Curva de característica del ítem/examen c. Correlaciones inter-ítem d. Correlaciones ítem-total 2. Confiabilidad de la escala de puntuación 3. Error estándar de la medida a. Generalizabilidad b. Dimensionalidad c. Análisis de factor del ítem d. Funcionamiento diferencial del ítem e. Modelo psicométrico 4. Ítems y opciones de respuesta (preguntas de opción múltiple, falso/verdadero) 5. Formato de la observación 6. Estandarización entre formatos y ocasiones 7. Rúbrica o criterio de puntuación, procedimientos de implementación, estándar de aprobado/no aprobado 8. Selección y entrenamiento de los evaluadores (p ej., ECOE) 9. Reglas para combinar los elementos relacionados con la prueba a partir de fuentes diferentes o para separar elementos no relacionados de la misma fuente 10. Seguridad de los datos y control de calidad	1. Análisis de dificultad de los ítems, correlaciones inter ítem, etc 2. Análisis de discernimiento de la pregunta 3. Análisis de las opciones de la pregunta 4. Análisis comparativo de grupos de candidatos 5. Análisis de confiabilidad. 6. Demostrar cuál es el porcentaje de calificación final que en realidad aporta cada componente. 7. Especificar cómo se deben combinar los puntajes de los componentes de la prueba y por qué. 8. Generalizabilidad, modelo psicométrico. 9. Desempeño de ítems y de opciones de respuesta 10. Formato de observación 11. Estandarización 12. Rúbrica o criterio de puntuación 13. Selección y entrenamiento de los evaluadores, confiabilidad y precisión de los evaluadores (p ej. en evaluación de desempeño – ECOE) 14. Seguridad de los datos y control de calidad

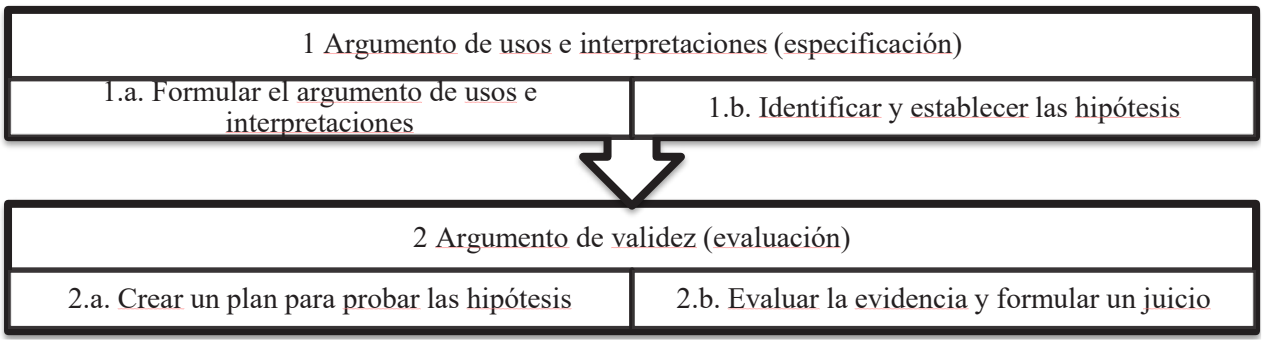
Generalización Extrapolación	Relación con otras variables	1. Correlación con otras variables relevantes 2. Correlación convergente interna y externa 3. Evidencia de generalizabilidad 4. Estrategia de muestreo de los ítems 5. Tamaño de la muestra (número de preguntas) 6. Alcance de la prueba 7. Autenticidad del contexto de la prueba 8. Autenticidad del ítem/escenario 9. Análisis que demuestren la relación entre el desempeño en la prueba y los dominios o contextos diferentes a los que se desea extrapolar	1. Correlación con las calificaciones de otras pruebas que evalúen constructos semejantes 2. Descripción general de cómo se obtiene la calificación de las otras pruebas con las que se busca la correlación. 3. Ecuaciones de regresión para obtener información acerca de predicción. 4. Medidas de tendencia central y medidas de dispersión. 5. Distribución del desempeño del criterio con respecto a la puntuación de la prueba en cuestión. Por rangos. 6. Correlación entre componentes que evalúan dominios semejantes 7. Confiabilidad o generalizabilidad por medio de la teoría de la generalizabilidad 8. Teoría de respuesta del ítem 9. Análisis para definir el alcance/objetivos 10. Acuerdo entre el proceso y el constructo 11. Relevancia y autenticidad 12. Correlación con otra medida que presente la misma relación esperada (con referencia al criterio o convergente; concurrente o predictiva) 13. Discriminación 14. Sensibilidad al cambio después de la intervención 15. Perfil del constructo 16. Funcionamiento diferencial del ítem
Implicaciones	Consecuencias	1. Impacto de los resultados de la prueba en los estudiantes y la sociedad 2. Consecuencias en el aprendizaje futuro 3. Evaluar si las consecuencias voluntarias sobrepasan a las involuntarias 4. Evaluar si es razonable el método para establecer aprobado/no aprobado 5. Estándar de aprobado/no aprobado 6. Acciones planeadas con base en los resultados de la prueba 7. Consecuencias voluntarias o involuntarias de las decisiones que se toman a partir de los resultados de la prueba.	1. Explicar y describir la cantidad de alumnos que presentan el examen y sus características poblacionales, así como sus resultados por componente y en general. 2. Rango de puntuaciones en cada componente de la prueba. 3. Considerar otros posibles objetivos de la prueba diferentes al establecido en los primeros pasos. 4. Explicar qué sucede si el sustentante aprueba o no aprueba el examen: con el sustentante, con su familia, con la sociedad 5. Explicar el tipo de escala que utiliza la prueba (por norma o por criterio). Si es por criterio, explicar y justificar cabalmente por qué se ha elegido ese criterio en particular para establecer el punto de aprobación. 6. Estándar de aprobado/no aprobado 7. Efectividad de las acciones basadas en los resultados de la prueba 8. Consecuencias voluntarias o involuntarias de la prueba 9. Funcionamiento diferencial del ítem

Figura 1. Puntos en común de los marcos de referencia de Messick y de Kane



Fuente: Elaboración propia.

Figura 2. Propuesta de método



Fuente: Basada en Tavares et al, (2018).

Referencias

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). Estándares para pruebas educativas y psicológicas. American Educational Research Association.

Brennan, R. (2013). Commentary on "Validating the Interpretations and Uses of Test Scores." Journal of Educational Measurement, 50(1), 74–83. https://doi.org/10.1111/jedm.12001

Clauser, B. E., Margolis, M. J., & Swanson, D. B. (2008). Issues of validity and reliability for assessments in medical education. In E. Holmboe & R. Hawkins (Eds.), A Practical Guide to the Evaluation of Clinical Competence (pp. 10–23). Elsevier.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. Medical Education, 49(6), 560–575. https://doi.org/10.1111/medu.12678

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281–302. https://doi.org/10.1037/h0040957

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. Medical Education, 37(9), 830–837. https://doi.org/10.1046/j.1365-2923.2003.01594.x

Embretson, S. E. (1998). A Cognitive Design System Approach to Generating Valid Tests: Application to Abstract Reasoning. Psychological Methods, 3(3), 380–396. https://doi.org/10.1037/1082-989X.3.3.380

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. Psychometrika, 64(4), 407–433. https://doi.org/10.1007/BF02294564

Hatala, R., Cook, D. A., Brydges, R., & Hawkins, R. (2015). Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. Advances in Health Sciences Education, 20(5), 1149–1175. https://doi.org/10.1007/s10459-015-9593-1

Kane, M. T. (1992). An argument-based approach to validity in evaluation. Psychological Bulletin, 112(3), 527–535. https://doi.org/10.1177/1356389011410522

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement, 50(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2016). Explicating validity. Assessment in Education: Principles, Policy and Practice, 23(2), 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (pp. 13–103). Macmillan. https://doi.org/10.1002/j.2330-8516.1987.tb00244.x

Messick, S. (1990). Validity of test interpretation and use. In ETS Research Report Series. https://doi.org/10.1002/j.2333-8504.1990.tb01343.x

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2012). Programmatic assessment and Kane's validity perspective. Medical Education, 46(1), 38–48. https://doi.org/10.1111/j.1365-2923.2011.04098.x

Tavares, W., Brydges, R., Myre, P., Prpic, J., Turner, L., Yelle, R., & Huiskamp, M. (2018). Applying Kane's validity framework to a simulation based assessment of clinical competence. Advances in Health Sciences Education, 23(2), 323–338. https://doi.org/10.1007/s10459-017-9800-3