ID: 10.33881/2027-1786.RIP.14102

ĬbērAM
Reflexión/Discusión | Reflection

Validez y Validación para Pruebas Educativas y Psicológicas

Teoria y recomendaciones

Validity and validation in educational and psychological testing: Theory and recommendations

Angel **Arias**
Stephen G. **Sireci**

Photo By/Foto: Lemau Studio

Rip 14(1)
Volumen 14 #1 ene-abr
14 Años

Revista Iberoamericana de
Psicología
ISSN-l: 2027-1786 | e-ISSN: 2500-6517
Publicación Cuatrimestral

Revista Iberoamericana de
Psicología
ISSN-l: 2027-1786 | e-ISSN: 2500-6517
Publicación Cuatrimestral

ID: 10.33881/2027-1786.RIP.14102

Title: Validity and validation in educatio and psychological testing:

Subtitle: Theory and recommendations

Título: Validez y Validación para Pruebas Educativas y Psicológicas

Subtítulo: Teoria y recomendaciones

Alt Title / Título alternativo:

[en]: Validity and validation in educatio and psychological testing:

Author (s) / Autor (es):

Arias & Sireci

Keywords / Palabras Clave:

[en]: validity; validation; sources of validity evidence; testing consequences; testing standards

[es]: validez;validación; fuentes de evidencias de validez; consecuenc de las pruebas; estándares de pruebas.

Submited: 2020-07-28

Acepted: 2020-10-14

Resumen

La validez es uno de los conceptos más fundamentales en el contexto de pruebas educativas y psicológicas y se refiere al grado en el que la evidencia teórica y empírica respaldan las interpretaciones de las puntaciones obtenidas a partir de una prueba utilizada para un fin determinado. El presente trabajo tiene como objetivo realizar una reflexión sobre recientes avances de la teoría de la validez y sugerir pautas a seguir para documentar evidencias necesarias para respaldar la interpretación de las puntuaciones de un instrumento de medida y su uso propuesto. Además, trazamos la historia de la teoría de la validez, centrándonos en su evolución y explicamos perspectivas y recomendaciones actuales para aplicar dicha teoría. Nos basamos en gran parte en los Estándares para Pruebas Educativas y Psicológicas, propuestos por la American Educational Research Association (AERA), la American Psychological Association (APA) y el National Council on Measurement in Education (NCME), los cuales proporcionan un marco conceptual para la validación de pruebas. También proporcionamos una breve descripción de la validación basada en argumentos y sus componentes, esbozando las dificultades asociadas a la operacionalización del proceso de validación desde una perspectiva de argumentación. Se proponen cinco fuentes de evidencias de validez de las puntuaciones de una prueba: contenido, procesos de respuesta, estructura interna, relaciones con otras variables y consecuencias. El uso de los Estándares permite que la evidencia de validez pueda ser acumulada de forma sistemática con el fin de respaldar la interpretación y el uso de las puntuaciones de una prueba para un propósito específico, promoviendo así prácticas solidas en cuanto al uso de un instrumento de medida lo cual puede contribuir a reducir las consecuencias negativas provenientes de la utilización de pruebas de alto riesgo.

Abstract

Validity is one of most fundamental concepts in the context of educational and psychological testing and refers to the degree to which theoretical and empirical evidence support the interpretations of test scores that are used for a particular purpose. The aim of this paper is to engage in a reflection of recent advances in validity theory and to suggest guidelines for documenting evidence needed to support the interpretation and uses of scores stemming from tests or other measurement procedures. In addition, we also trace the history of validity theory, focusing on its evolution, explicating current perspectives and recommendations on how to apply this theory. We draw heavily on the Standards for Educational and Psychological Testing, proposed by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), which provide a framework for test validation. We also provide a brief description of argument-based validation and its components, outlining the difficulties and challenges associated with operationalizing the validation process from an argumentation perspective. Five sources of validity evidence are proposed for the validation of test scores: test content, response processes, internal structure, relations to other variables, and consequences of testing. The use of the Standards allows validity evidence to be systematically accumulated and documented in order to support the interpretation and use of test scores for a specific purpose, thus promoting sound testing practices that can contribute to reducing the negative consequences of high-stakes tests.

Citar como:

Arias, A., & Sireci, S. G. (2021). Validez y Validación para Pruebas Educativas y Psicológicas . Revista Iberoamericana de Psicología , 14 (1), 11-22. Obtenido de: https://reviberopsicologia.ibero.edu.co/article/view/1926

Angel **Arias**, PhD. MA
ORCID: *https://orcid.org/0000-0001-8565-6030*

Source | Filiacion:
Carleton University

BIO:
holds a Ph.D. in Educational Measurement from the Université de Montréal, Canada; his master's degree is in Applied Linguistics and Discourse Studies from Carleton University, Canada and his bachelor's degree is in Education from Universidad Dominicana Organización y Métodos (O&M), Dominican Republic. He is Assistant Professor of Applied Linguistics in the School of Linguistics and Discourse Studies at Carleton University. His research interests focus on the application of psychometric models and mixed methods approaches in language testing and assessment to evaluate validity evidence of test score meaning and justification of test use in high stakes contexts and classroom assessment. He has served as an external consultant for the Ministry of Quebec's Education and Chair of the Test Validity Research and Evaluation special interest group of the American Educational Research Association (AERA). He speaks Spanish, English, and French fluently.

City | Ciudad:
Ottawa, ON, [ca]

Stephen G. **Sireci**, PhD. MA
ORCID: *http://orcid.org/0000-0002-2174-8777*

Source | Filiacion:
University of Massachusetts

BIO:
Distinguished University Professor and Director of the Center for Educational Assessment in the College of Education at the University of Massachusetts Amherst. He is known for his research in evaluating test fairness, particularly issues related to content validity, test bias, cross-lingual assessment, standard setting, and computerized-adaptive testing. He has authored/coauthored over 130 publications, and is the co-architect of the multistage-adaptive Massachusetts Adult Proficiency Tests. He is a Fellow of the American Educational Research Association, and of Division 5 of the American Psychological Association; Past-President of the National Council on Measurement in Education, and President-Elect of the International Test Commission.

City | Ciudad:
Amherst, MA, [us]

Validez y Validación para Pruebas Educativas y Psicológicas

Teoria y recomendaciones

Validity and validation in educational and psychological testing: Theory and recommendations

Angel **Arias**
Stephen G. **Sireci**

El diccionario virtual de la Real Academia Española (2020) define validez como “cualidad de válido” y el término válido es definido como “firme, subsistente y que vale o debe valer legalmente”. Cuando se trata de la evaluación educativa o psicológica, esta última definición es probablemente la forma en la que el público general piensa del concepto de validez – una cualidad que una prueba o medida estandardizada posee o no posee. En los inicios del testing, la noción de que la validez era inherente a una prueba era algo común. Sin embargo, la conceptualización contemporánea de validez enfatiza que la validez no es una propiedad de una prueba, sino una cualidad que se refiere a cómo se interpretan y se usan las puntuaciones de las pruebas. Por esta razón, los Estándares para Pruebas Educativas y Psicológicas, desarrollados por la American Research Association (AERA), la American Psychological Association (APA) y el National Council on Measurement in Education (NCME) definen la validez como “…el grado en que la evidencia y la teoría respaldan las interpretaciones de las puntuaciones de una prueba para usos propuestos de las pruebas” (AERA, APA, y NCME, 2014/2018, p. 11). Por lo tanto, en el caso de pruebas educativas e instrumentos de medida en psicología, la pregunta sobre la validez no es “¿es este instrumento de medida válido?”, sino “¿son las interpretaciones hechas a partir de las puntuaciones obtenidas válidas con respecto a un uso en particular? Esta distinción puede parecer trivial, pero es importante señalar que dicha distinción caracteriza el concepto de validez en contextos muy específicos basados en la manera en que las puntuaciones se interpretan y se utilizan en un momento dado.

Revista Iberoamericana de **Psicología**
ISSN-L: 2027-1786

Un concepto relacionado con validez es validación y puede definirse como el proceso mediante el cual se acumulan las evidencias para respaldar las interpretaciones y los usos específicos de las puntuaciones de las pruebas. La idea de validar la interpretación de las puntaciones y los usos de una prueba, en lugar de validar una prueba, es fundamental para comprender las nociones de validez del siglo XXI. La validación del uso de las pruebas crea la necesidad de definir claramente su propósito y vincula las interpretaciones de las puntuaciones estrictamente al uso propuesto. Lo cual ayuda a evitar la práctica de reorientación, la cual consiste en usar las pruebas para fines no previstos (por ejemplo, utilizar una prueba originalmente diseñada para admisión a programas de estudios universitarios para tomar decisiones de empleos que no están relacionados con el constructo de la prueba).

La validación del uso apropiado de una prueba requiere que se preste atención a las consecuencias de esta. Es decir, el proceso de validación debe considerar el impacto que una prueba tiene en las personas y en la sociedad. Naturalmente, algunas pruebas tienen consecuencias mucho más importantes que otras, por lo que la intensidad del proceso de validación variará en consecuencia. Mientras más ambiciosas sean las alegaciones o afirmaciones a favor del uso de una prueba, mucho más se necesitarán evidencias que las respalden **(Kane, 2013)**. La evidencia que respalda el uso de una prueba para un propósito en particular será diferente dependiendo de los individuos a los que se pretende evaluar, su propósito, la forma en que se interpretarán las puntuaciones, y otros factores. Por ejemplo, una prueba que se utiliza para certificar a cirujanos como aptos de operar a personas requerirá tipos de evidencia que podrían no requerirse para una prueba que se utiliza para evaluar el conocimiento de adición en preescolar.

En este trabajo, presentamos los conceptos de validez y validación basándonos principalmente en los Estándares para Pruebas Educativas y Psicológicas de la AERA et al. (2014/2018). Es importante destacar que las tres organizaciones que desarrollaron los Estándares (AERA, APA y NCME) han estado colaborando durante más de 60 años para "proporcionar criterios para el desarrollo y la evaluación de pruebas y prácticas de desarrollo de pruebas y brindar pautas para evaluar la validez de las interpretaciones de las puntuaciones de las pruebas para los usos previstos de las pruebas" (p. 1). Por lo tanto, recomendamos que los lectores consulten los Estándares de la AERA et al. **(2014)** para una discusión más completa de prácticas sólidas y rigurosas para el desarrollo y validación de pruebas, ya que se les considera ampliamente como una fuente de referencia muy importante **(Plake y Wise, 2014; Sireci, 2016; Sireci y Parker, 2006)**.

Gran parte de la literatura sobre la teoría de la validez ha sido escrita en inglés y los Estándares han sido elaborados en esta lengua, con la excepción de la versión más reciente **(AERA et al., 2014)**, la cual fue traducida al español (véase *https://www.aera.net/Publications/-Online-Store/Books-Publications/BKctl/ViewDetails/SKU/AERSTDSPMAIN*) y la versión de 1999 la cual fue traducida al francés **(Sarrazin, 2003)**. El objetivo central del presente trabajo es de presentar los avances más recientes de la teoría de la validez y las prácticas de validación a la comunidad hispanohablante y proporcionar pautas para guiar la validación de pruebas o instrumento de medida. Como ya hemos mencionado, nos basamos en gran parte en los Estándares ya que estos se consideran discutiblemente la posición consensuada y contemporánea de la teoría de la validez. Estos constituyen la corriente dominante del tema, son muy influyentes y tienen un gran potencial para ser aplicados a nivel internacional **(Zumbo, 2014)**, especialmente en contextos de pruebas cuyos usos pueden tener serias repercusiones en la vida de las personas y la sociedad. Además, la mayoría de los artículos sobre la teoría de la validez producidos por la comunidad científica ofrecen mayormente perspectivas del tema basándose en los Estándares o haciendo referencia a estos. Iniciamos con una breve historia del concepto de validez. Luego, presentamos diferentes fuentes de evidencia de validez, y cómo pueden ser integradas en un argumento de validez **(Kane, 1992, 2006, 2013)** para respaldar el uso de una prueba para un propósito determinado.

Breve historia del concepto de validez

La validez se considera como uno de los conceptos más fundamentales de la psicometría **(Sireci, 2009)** y ha evolucionado considerablemente desde el modelo tripartito de contenido, de predicción y de constructo **(APA, 1966)**. Sireci (2009, 2016) describe dos conceptualizaciones históricas de validez que aún persisten. La primera es la noción de que la validez se refiere al "grado en que una prueba mide lo que pretende medir". Esta definición se remonta a los investigadores pioneros en medición en educación (psicómetras) que utilizaron el análisis factorial para identificar "rasgos" hipotéticos que las pruebas supuestamente medían **(véase, Garrett, 1937; Smith y Wright, 1928).** La segunda definición define la validez en términos de correlaciones de los resultados de las pruebas, y propone que "una prueba es válida con todo lo que se correlacione con ella" **(Guilford, 1946, p. 429)**. Esta definición se considera simplista en la actualidad, pero como se discutirá más adelante, la evaluación de las correlaciones de las puntuaciones de las pruebas con otras variables sigue siendo un área importante en el proceso de validación.

Para que una prueba sea considerada útil para un propósito determinado, se necesita evidencias que respalden que la prueba mide lo que pretende medir. Por lo tanto, la evaluación de dicha afirmación es una parte importante de la validación. Sin embargo, debido a que el uso de las puntuaciones de las pruebas siempre implica algún tipo de acción o consecuencia (por ejemplo, certificación, remediación, admisión a la universidad, etc.), la simple demostración de que la prueba mide el conocimiento o las áreas de competencia que se diseñó para medir, no es suficiente para defender su uso. Por esta razón, la noción de que la validez se refiere al grado en que una prueba mide lo que pretende medir se ha considerado como una definición incompleta por más de 70 años. Por ejemplo, Rulon **(1946, p.290)** comenta que "esta es una noción insatisfactoria de validez y poco útil, ya que, en este sentido, la validez de una prueba puede verse alterada por completo si se cambia arbitrariamente su propósito".

Hasta ahora hemos tratado tres definiciones de validez. Dos son de interés histórico y son obsoletas. La tercera y más importante definición es la que proporcionan los Estándares de la AERA et al. (2014) y amerita ser repetida: "la validez se refiere al grado en que la evidencia y la teoría respaldan las interpretaciones de las puntuaciones de una prueba para usos propuestos de las pruebas" (p. 11).

En la Tabla 1 se presentan las diferentes formas en que se ha definido la validez en cada versión de los Estándares. La primera columna presenta el título de cada versión de lo que hoy conocemos como los Estándares. La segunda columna muestra la terminología utilizada para describir el concepto de validez. Varios términos en esta columna se pueden encontrar aún en libros modernos de medición, tales como validez de constructo, validez de contenido, validez predictiva y validez concurrente.

Debido a que la validez es un concepto unitario, los términos validez de constructo, validez de contenido y validez predictiva no se utilizan como "tipos" de validez, sino que estos procesos subsisten dentro

de las cinco fuentes de evidencia de validez de los Estándares de la AERA et al. (2014). En las secciones siguientes se explican estos conceptos y su relación con los Estándares vigentes.

Validez de constructo, de contenido y de criterio

El término "constructo" fue introducido por Cronbach y Meehl (1955) y ha sido objeto de exhaustivos diálogos académicos en psicología **(véase, Lovasz y Slaney, 2013; Slaney y Racine, 2013)**. En términos simples, un constructo se refiere a la definición de "algún atributo de las personas, que se supone reflejarse en el rendimiento en las pruebas" (**Cronbach y Meeh, 1955, p. 283)**. Esencialmente, el conocimiento, la competencia, la habilidad u otro atributo medido por una prueba o encuesta se considera como un constructo. Los constructos medidos por pruebas educativas y psicológicas no pueden ser observados directamente, por lo que son considerados como constructos subyacentes o rasgos hipotéticos. La validez de constructo emergió como un tipo de validez que describía el grado en que la prueba medía el constructo propuesto. Inicialmente se utilizaron evidencias de análisis factorial para proveer evidencia de constructo, pero Loevinger (1957), Messick (1989) y otros, argumentaron que toda validación era validación de constructo.

El término validez de constructo ya no se utiliza en los Estándares porque es sinónimo de validez en general. Sin embargo, sigue siendo teóricamente relevante para la validación porque al evaluar la calidad de una prueba, los aspectos que son pertinentes para el constructo de interés, y los que no lo son, pueden ser objeto de un estudio de validez. Como señaló Messick (1989), "las pruebas son medidas imperfectas de los constructos porque no incluyen todo lo que debería incluirse o bien incluyen algo que no debería incluirse, o ambas cosas" (p. 34). Estos problemas se conocen respectivamente como infrarrepresentación de constructo y varianza irrelevante de constructo.

Los términos validez de contenido, validez concurrente, validez de criterio y validez predictiva ya no se utilizan, pero las nociones a las que aluden están contenidas en las evidencias de validez que se han utilizado para describir el concepto de validez desde la versión de 1999 de los Estándares. Por ejemplo, la validez de criterio se refiere al grado en el que las puntuaciones de la prueba se correlacionan con otras variables relacionadas con el constructo que la prueba pretende medir. Las "otras variables" se denominan "criterios". La validez predictiva y la validez concurrente son subcategorías de la validez de criterio y se refieren al momento en que se obtienen los datos del criterio. La validez predictiva se refiere a los datos de los criterios obtenidos con posterioridad a la obtención de los resultados de las pruebas (por ejemplo, las calificaciones universitarias utilizadas como criterio de validación para los resultados de las pruebas de escolaridad previa). La validez concurrente se refiere a los datos del criterio obtenidos prácticamente al mismo tiempo que los resultados de la prueba (por ejemplo, el uso de una forma extensa de una prueba como criterio de una forma corta de la misma). El valor de este tipo de evidencia está contenido en las fuentes de validez de los Estándares de la AERA et al. y, por lo tanto, en lugar de detallarlas en esta sección, serán abordadas en la sección sobre las fuentes de evidencia de validez que han sido establecidas en los Estándares desde 1999.

Si bien los términos de validez de contenido, validez concurrente, validez de criterio y validez predictiva se consideran una terminología obsoleta, varias investigaciones que documentan evidencias de validez continúan utilizando estos términos (por ejemplo: Aguirre, 2014; Bermúdez, 2010; Díaz et al., 2013; Riveros et al., 2015; Ronquillo et al., 2013) y excesivamente se reducen a trabajos de evidencia de constructo utilizando la modelación de análisis factorial (Sakakibara et al., 2020). Las evidencias de constructo (es decir, la estructura interna) son de suma importancia, pero la idea de que toda validación es considerada validación de constructo no sugiere que la evidencia de la estructura interna de una prueba o encuesta sea suficiente para establecer la validez de las interpretaciones y el uso de una prueba o instrumento de medida.

Antes de finalizar este breve recorrido histórico de la literatura sobre la validez, es importante destacar los trabajos de Mislevy (2009, 2018), quien ha extendido las consideraciones sociales de Messick (1989) con respecto al uso de las pruebas enmarcando la validación dentro de una perspectiva "sociocognitiva", explicando cómo las diferencias en la modelización psicométrica de una prueba implican y requieren diferentes conexiones formales entre el modelo y las inferencias derivadas de las puntuaciones de las pruebas. Según explica Mislevy (2009), "un elemento esencial de la validez de una prueba es si, en una aplicación determinada, la utilización de un modelo determinado proporciona una base sólida para organizar las observaciones y orientar las acciones en las situaciones para las que está destinada" (p. 83). Esta perspectiva de validez enfatiza el uso de las pruebas, al igual que los estándares de AERA et al (1999, 2014); y al igual que Messick (1989), también reconoce las variaciones del contexto social en el que se produce la evaluación.

Si utilizamos términos de la filosofía de la ciencia, la perspectiva sociocognitiva de Mislevy favorece "una visión constructivista-realista de la validez**" (Mislevy, 2009, p. 84)**, que también fue abordada por Messick (1989). La perspectiva es "realista" en el sentido de que supone la existencia real del constructo que se pretende evaluar, pero "constructivista" en el sentido de que se reconoce que el constructo que se pretende evaluar puede variar ampliamente en función de los responsables del desarrollo de la prueba, las condiciones de medición y el contexto. Como describe Mislevy (2009), "La visión constructivista-realista sostiene que los modelos son construcciones humanas, pero los modelos eficaces disciernen y expresan las estructuras que caracterizan los aspectos de los fenómenos más complejos del mundo real" (p. 95). Por lo tanto, la perspectiva sociocognitiva es congruente con la definición de los Estándares en el sentido de que la validez hace referencia al grado en que el uso de una prueba para un fin determinado es justificado mediante pruebas y teoría, y por consiguiente es más útil que conceptualizaciones restrictivas que provienen de una perspectiva puramente realista (por ejemplo, Borsboom, Mellenbergh, y van Heerden, 2004).

Cinco fuentes de evidencia de validez

Los Estándares de la AERA et al. (2014) describen cinco "fuentes de evidencia que podrían utilizarse en la evaluación de la validez de una interpretación propuesta de las puntuaciones de una prueba para un uso particular" (p. 14). Estas cinco fuentes de evidencia de validez son basadas en (a) el contenido de la prueba, (b) los procesos de respuesta, (c) la estructura interna, (d) las relaciones con otras variables, y (e) las consecuencias de las pruebas. Los Estándares enfatizan que las cinco fuentes de evidencia de validez no son distintos tipos de validez, sino que contribuyen al conjunto de evidencia que puede ser utilizada para respaldar el uso de una prueba para un propósito determinado. A continuación, describimos cada una de estas cinco fuentes de evidencia de validez, proveyendo ejemplos de metodologías para recolectar cada tipo de evidencia. Las metodologías y técnicas de análisis incluidas en este trabajo las utilizamos como ejemplos y no

RIP 14 (1) pág. 11 -22
Revista Iberoamericana de Psicología
ISSN-L: 2027-1786

como una prescripción, ya que existen varios métodos que se pueden utilizar para acumular el tipo de evidencias necesarias para respaldar cada una de las evidencias de validez descritas en los Estándares.

Evidencia de validez basada en el contenido de la prueba

La evidencia de validez basada en el contenido de la prueba se refiere a la evidencia que se utiliza para evaluar el grado en que el contenido de la prueba es consistente con el propósito de la prueba y representa suficientemente los conocimientos, las competencias, o las habilidades que se desean medir. La evidencia de contenido es generalmente recolectada a través de expertos en el área de evaluación con el fin de evaluar los ítems y calificarlos con respecto a su relevancia con el dominio que se quiere evaluar **(Crocker, Miller, y Franks, 1998; Sireci, 1998; Sireci y Faulkner-Bond, 2014)**. Los estudios de alineamiento o alineación **(Anderson et al., 2015; Bhola et al., 2003; Cizek et al., 2018; Martone y Sireci, 2009; Russell y Moncaleano, 2020; Traynor, 2017; Webb, 2007)** son de suma importancia para recopilar evidencia sobre la congruencia de los ítems y el dominio que se desea evaluar. La evidencia de contenido debe garantizar que la definición del contenido y los dominios cognitivos evaluados sean apropiados para el propósito de la prueba y que los ítems representan suficientemente el dominio evaluado **(Sireci y Faulkner-Bond, 2014)**. El uso de evidence-centered design y assessment engineering como enfoques de elaboración de pruebas ha adquirido bastante atención y es instrumental para documentar evidencias basadas en el contenido de una prueba o instrumento de medida **(Luecht, 2013; Mislevy et al., 2003; Risconscente et al., 2015)**.

Evidencia de validez basada en los procesos de respuesta

La evidencia de validez basada de los procesos de respuesta incluye aspectos relacionados a la forma en que las personas evaluadas interactúan con una prueba. Específicamente, se necesitan evidencias empíricas para confirmar que las personas evaluadas movilizan los procesos cognitivos previstos al responder a los ítems **(Ercikan y Pellegrino, 2017; Zumbo y Hubley, 2017)**. La recopilación de tales evidencias implica indagar sobre los procesos cognitivos utilizados durante la prueba. Los tipos de evidencias en esta área incluyen la práctica de protocolo verbal mientras las personas evaluadas responden a los ítems **(Leighton, 2004)**, la realización de entrevistas cognitivas **(Padilla y Benítez, 2014)**, la evaluación de las pulsaciones de teclado y el seguimiento de los movimientos oculares **(Bax, 2013)** o la medición del tiempo de reacción ante los ítems **(van der Linden, 2009)**. La evidencia de procesos de respuesta es particularmente importante para las pruebas que afirman o pretenden medir las habilidades de pensamiento de alto nivel. Es posible que las personas evaluadas puedan resolver los ítems usando atajos o estrategias que no requieran el uso de las habilidades cognitivas previstas, como, por ejemplo, resolver un problema matemático insertando las opciones de un ítem de selección múltiple en una fórmula dada en la pregunta del ítem, en lugar de resolver el problema con los datos facilitados. Del mismo modo, es posible responder a preguntas de comprensión de lectura usando únicamente la pregunta del ítem, sin leer el pasaje o el poema que lo acompaña. Solo podemos confirmar que realmente se están midiendo las habilidades de nivel superior cuando se investiga que en realidad éstas son movilizadas al responder a los ítems.

Evidencia de validez basada en la estructura interna

La evidencia de validez basada en la estructura interna incluye varias áreas técnicas para evaluar la calidad de las puntuaciones de las pruebas y pueden incluir desde la precisión y confiabilidad de esta hasta su dimensionalidad. Con respecto a la dimensionalidad, dicha evidencia podría confirmar la unidimensionalidad de una prueba lo cual es necesario para respaldar el uso de un modelo específico de puntuación. Además, la dimensionalidad observada podría utilizarse para evaluar si los datos son consistentes con la teoría que respalda el constructo y las puntuaciones de la prueba. Por ejemplo, una prueba lingüística puede especificar cuatro dominios o dimensiones: expresión oral, expresión escrita, comprensión auditiva y comprensión escrita. Si un análisis factorial (confirmatorio o exploratorio) determinara que los ítems que miden cada uno de estos dominios forman dimensiones distintas (factores), entonces se sustentaría la teoría que respalda la prueba. Se recomienda que cuando se aplican modelos de análisis factorial, se verifiquen los índices de bondad de ajuste para sustentar la solución obtenida y obtener resultados robustos. Las líneas directrices recomiendan un valor de Chi-cuadrado no significativo – pero tómese en cuenta la sensibilidad de este índice al tamaño muestral. Conjuntamente, se deben considerar otros índices de ajuste tales como el ajuste comparativo (CFI), el ajuste no normado (TLI) cuyos valores aceptables a menudo se fijan en y el error cuadrático medio de aproximación (RMSEA) con valor aceptable de **(Hu y Bentler, 1999)**. Estas recomendaciones no son fijas, ni implican una receta al pie de la letra, al contrario, se recomienda tomar en consideración el contexto en el cual se utiliza la prueba. Existen recomendaciones clásicas para el análisis factorial (exploratorio o confirmatorio), pero ya estas recomendaciones han sido revisadas y se recomienda consultar las recomendaciones actuales **(Lloret-Segura et al., 2014)** para una implementación óptima de esta modelación psicométrica.

Con respecto a la confiabilidad o precisión de la prueba, las evidencias de la estructura interna podrían incluir funciones de información de la teoría de respuesta a los ítems y errores estándar condicionales que indiquen la precisión de una medida en la escala de puntuaciones de la prueba. Cuando se utiliza un coeficiente de confiabilidad, se recomienda utilizar alternativas al coeficiente Alfa, debido a las limitaciones asociadas a este índice. Estas limitaciones incluyen la cantidad de ítems, el número de opciones de respuesta, la proporción de la varianza y la dimensionalidad de la prueba **(Ventura-León y Caycho-Rodríguez 2017)**. Nótese que es posible estimar apropiadamente la confiabilidad de las puntuaciones de un instrumento de medida mediante el coeficiente Alfa, pero para ello, se espera que los ítems sean al menos tau-equivalentes de lo contrario existe la posibilidad de que la estimación sea sesgada **(Graham, 2006)**. Debido a los estrictos supuestos del coeficiente Alfa, regularmente se recomienda usar el coeficiente Omega como una alternativa menos afectada por las limitaciones que afectan el coeficiente Alfa y por ende se considera un índice con un mínimo nivel de sesgo. Sin embargo, debido a que la parametrización del coeficiente Omega se obtiene por medio de análisis factoriales, es de suma importancia que la modelación refleje las características de los datos para así elegir la variante más adecuada del coeficiente Omega. Por ejemplo, si los datos son multidimensionales se debe especificar un modelo de análisis factorial multidimensional para poder estimar adecuadamente el coeficiente Omega. Además, se debe tomar en cuenta el tipo de datos (continuos o categóricos) ya que estos requieren métodos de estimación diferentes. La calidad del coeficiente Omega depende enormemente de la correcta aplicación de modelos factoriales ya que se debe elegir la mejor alternativa de dicho

coeficiente (unifactorial, multifactorial, jerárquico-factorial, bifactorial, etc.), respondiendo a las características de los datos para los cuales se estima este coeficiente **(Flora, 2020)**.

Otros tipos de evidencias de estructura interna incluyen estudios de funcionamiento diferencial de los ítems, en los que se evalúa la invariabilidad del instrumento de medida entre grupos específicos. Estos estudios son un paso preliminar en la investigación de sesgo de los ítems. Además del funcionamiento diferencial de los ítems, es conveniente y recomendable investigar la invarianza de la prueba, el cual no se restringe únicamente a la invariabilidad de los ítems para los distintos grupos, sino también a la invariabilidad del número de factores y la cantidad de ítems por factor, los valores de las cargas de los factores, los errores de medición y la varianza o invarianza de las variables latentes **(Wells, 2021)**. Este tipo de investigación es particularmente necesaria cuando las pruebas o encuestas son traducidas o adaptadas **(Internacional Test Commission [ITC], 2017)**; ya que es necesario acumular evidencia de invariabilidad de constructo después de que la traducción se ha llevado a cabo **(Sireci et al., 2016)**.

Evidencias de validez basada en relaciones con otras variables

La evidencia de validez basada en relaciones con otras variables se refiere a la evidencia de los estudios que analizan las puntuaciones de las pruebas y otras variables relacionadas con el constructo de interés. Por lo general, estos estudios utilizan técnicas de regresión o correlación para examinar las relaciones de prueba-criterio. Campbell y Fiske (1956) distinguen entre los constructos que se espera que estén relacionados con las puntuaciones de la prueba (evidencia convergente), y los constructos que no deberían tener ninguna relación con las puntuaciones de la prueba (evidencia discriminante); este tipo de evidencia es regularmente recopilada a través de estudios de matrices multirasgo-multimétodo. Otros tipos de estudios que podría movilizarse para documentar evidencias de validez basada en relaciones con otras variables incluyen estudios longitudinales, análisis de trayectoria y análisis de ecuaciones estructurales **(Byrne, 2014)**.

Los estudios de evidencia predictiva y concurrente descritos anteriormente también están dentro de esta categoría, al igual que algunos estudios que investigan las diferencias de puntuaciones entre grupos. Por ejemplo, los estudiantes pueden ser asignados aleatoriamente a grupos de instrucción o de control y luego recibir un examen después de que la instrucción haya ocurrido. Si el examen es válido para medir la eficiencia de la instrucción, los estudiantes en el grupo de control deberían obtener resultados menos satisfactorios. En este caso, la variable de agrupación (grupo de instrucción o de control) es la variable externa con la que se relacionan las puntuaciones de la prueba.

Es importante enfatizar que cuando se realizan estudios de correlación, es de suma importancia verificar que los supuestos del análisis realizado sean verificados y utilizar la técnica apropiada de acuerdo con los tipos datos obtenidos. El uso inadecuado de una correlación paramétrica o no paramétrica puede artificialmente aumentar o disminuir el coeficiente de correlación. De la misma forma, el tipo de datos obtenido puede requerir el uso de coeficientes correlacionales menos comunes tales como los coeficientes correlacionales tetracórico, policórico, poliserial, biserial, etc.

Otro concepto importante para la evidencia de relación con otras variables es la generalización de la validez, que es una técnica meta-analítica que se puede utilizar para explorar o determinar si los resultados de estudios individuales pueden sintetizarse y luego generalizarse a otros contextos. La generalización de la validez es más factible que se utilice cuando se ha llevado a cabo un número sustancial de estudios de validez en el contexto en el que se basa el metaanálisis.

Evidencia de validez y consecuencias de las pruebas

Las evidencias de validez basadas en las consecuencias de las pruebas se refieren al análisis de las repercusiones de las pruebas en las personas y en la sociedad. En cierto sentido, esta es la categoría más general de validez porque las pruebas están diseñadas para tener consecuencias específicas, es decir, están diseñadas para propósitos específicos y para tomar decisiones. Por lo tanto, la evaluación de las consecuencias de una prueba consiste en determinar en parte si los resultados de una prueba satisfacen su propósito y, al mismo tiempo, que no tengan ningún efecto negativo. El análisis de las consecuencias de las pruebas se inscribe en la filosofía de Cronbach (1989) y Messick (1989), quienes sostienen que las consecuencias del uso de las pruebas deben ser una de las principales preocupaciones de cualquier trabajo de validación.

Messick (1989) destaca que las consecuencias de las pruebas pueden ser positivas o negativas, y pueden ser previstas o no. Sin duda, las consecuencias negativas no son intencionadas, pero ya que pueden ocurrir, se recomienda a los responsables de documentar evidencias de validez que consideren cómo los resultados de las pruebas pueden ser malinterpretados o mal utilizados. Las consecuencias negativas no deseadas o imprevistas pueden tener un impacto adverso cuando los subgrupos evaluados obtienen índices de aprobación o de selección sustancialmente distintos. Otra consecuencia negativa puede ser que los profesores reduzcan su plan de enseñanza a "enseñar para la prueba", a expensas de objetivos curriculares más importantes. Además, si el uso de una encuesta o instrumento de medida incita a que las personas respondan de manera sesgada, las consecuencias del uso de la encuesta podrían ser negativas.

Entre los métodos para obtener evidencias de validez basadas en las consecuencias de las pruebas se incluyen encuestas a las personas implicadas en el proceso de evaluación, tales como, estudiantes, padres o profesores o analizando los cambios en las prácticas de enseñanza mediante estudios de impacto **(Lane, 2014)**. Puede ser difícil de obtener evidencias de consecuencias ya que estas pueden ocurrir mucho después de pasar la prueba y pueden ser imperceptibles. Por esta razón, recomendamos a los responsables del proceso de evaluación que consideren el uso inadecuado y la incorrecta interpretación de los resultados de las pruebas en las etapas iniciales de su desarrollo, y que evalúen si tales problemas ocurren durante la vigencia de un programa de evaluación educativo o psicológico. Los Estándares de la AERA et al. (2014) indican de forma explícita que las consecuencias de las pruebas pueden ser positivas o negativas, y que cuando son positivas, los efectos positivos pueden ser utilizados como evidencia para respaldar la interpretación de las puntuaciones para un uso determinado. Sin embargo, si se observan consecuencias negativas, éstas pueden ser indicio de un problema en cuanto al uso de las puntuaciones para el propósito previsto.

Otro aspecto importante y estrechamente relacionado a las consecuencias del uso de las pruebas es el concepto de justicia, el cual es importante considerar cuando las consecuencias del uso de las pruebas tienen un gran impacto sobre la sociedad e individuos **(Dorans y Cook, 2016; Jonson et al., 2019)**. El concepto de justicia en evaluación educativa y psicológica es bastante amplio y no existe una definición consensual. Sin embargo, se recomienda considerar el sistema de valores que subyace la interpretación de justicia en un contexto de-

RIP 14 (1) pág. 11 -22
Revista Iberoamericana de Psicología
ISSN-L: 2027-1786

terminado y reflexionar sobre las decisiones basadas en mérito y en equidad, ya que pueden informar substancialmente reflexiones sobre la justicia de una prueba o instrumento de medida **(Boyer, 2020)**. El concepto de justicia también ocupa un lugar importante en los Estándares, los cuales dedican un capítulo completo sobre el tema, y se sugiere considerarse en conversaciones sobre la validez y el uso de pruebas. Esto significa diseñar y desarrollar pruebas que sean justas para todos los estudiantes o personas que responden a una prueba o encuesta sin importar sus características demográficas o culturales **(Kunnan, 2010)**, incluyendo personas con necesidades especiales y diversas discapacidades **(Randall y García, 2016; Sireci et al., 2018)**. Esto implica proporcionar un campo de juego equitativo para todos y todas para promover un mejor rendimiento escolar y erradicar obstáculos que puedan contaminar los datos que provienen de las pruebas o instrumentos de medida **(Sireci, 2020)**.

Las fuentes de evidencias de validez propuestas en los Estándares son de suma importancia y ofrecen pautas para recolectar la información necesaria para respaldar las interpretaciones de las puntuaciones y el uso de una prueba en un contexto determinado. Cada fuente de evidencia de validez desempeña un rol específico, por lo tanto, los tipos de estudios que podrían realizarse o el tipo de información que se debería recopilar para construir un argumento de validez están estrechamente relacionado con las afirmaciones que se avanzan a favor del uso de una prueba o instrumento de medida. La información sobre el tipo de estudio que se puede movilizar para cada una de las fuentes de validez no debe considerarse como prescriptiva o exhaustiva, ya que existen formas de obtener evidencias para sustentar cada una de las fuentes de validez propuestas en los Estándares.

El concepto de validación basado en argumentos

Kane (1992) introdujo un modelo de la validación basado en argumentos que consiste en acumular evidencias exhaustivas y coherentes que respalden el uso de una prueba para un fin determinado **(véase también Kane, 2006, 2013)**. Este planteamiento es ampliamente aceptado porque reconoce que las pruebas nunca pueden ser "absolutamente" validadas para un propósito determinado ya que, para defender su uso para un propósito particular, se requiere un conjunto sustancial de evidencias basadas tanto en la teoría como en el estudio empírico. Los Estándares de la AERA et al. (2014) adoptaron esencialmente una perspectiva de argumentación porque describen que "un argumento de validez sólido integra diversos aspectos de la evidencia en una explicación coherente del grado en que la evidencia existente y la teoría respaldan la interpretación prevista de las puntuaciones de la prueba para usos específicos" (p. 23).

Chapelle (2020) se inspira de los trabajos de Kane (1992, 2006, 2013) y expone exhaustivamente la validación basada en argumentos, proveyendo ejemplos concretos de cómo operacionalizar la validación desde una perspectiva de argumentación **(Toulmin, 2003)** para pruebas existentes y pruebas nuevas. Estos trabajos utilizan términos de la teoría de la argumentación y los adoptan para facilitar la organización de argumentos de validez. Sin embargo, hay una serie de elementos de nivel macro, mezzo y micro que son indispensables para aplicar este enfoque. Por ejemplo, el proceso de validación inicia proponiendo una afirmación o inferencia para una interpretación o uso determinado de las puntuaciones de una prueba (por ejemplo, "las puntuaciones de la prueba son confiables"). Luego, para respaldar esta macro inferencia se necesitan otros detalles en forma de garantías, supuestos y respaldos para avalar la afirmación o inferencia. Las garantías pueden considerarse de un nivel de detalle mezzo, ya que tienden a ser menos generales que las afirmaciones o inferencias, pero no ofrecen suficientes detalles sobre cómo recopilar las pruebas empíricas necesarias para respaldar las interpretaciones de las puntuaciones y el uso determinado de una prueba. Los supuestos se despliegan a partir de las garantías y proporcionan un nivel micro de detalle que regularmente indica claramente los estudios que deberían realizarse para autorizar la interpretación de las puntuaciones y el uso propuestos de las pruebas, como tal, los estudios empíricos se consideran fundamentos o respaldos en la terminología de la validación basada en argumentos. Este efecto dominó o cadena acumulativa de conceptos estrechamente relacionados hace que la validación basada en argumentos sea una tarea difícil de aplicar. A pesar de estas dificultades y desafíos, la validación basada en argumentos ha alcanzado considerable notoriedad en diversos campos (por ejemplo, evaluación de lenguas y ciencias médicas) y sigue adoptándose como marco de referencia en el proceso de validación de pruebas. No obstante, se requiere un nivel adecuado de conocimientos sobre la validación basada en argumentos para aplicar este enfoque de manera coherente y sistemática. Chapelle (2020) hace una gran contribución al proporcionar una presentación ejemplar de la validación basada en argumentos, la cual tiene un gran potencial para formar a profesionales en cuanto a las pautas de cómo poner en práctica este enfoque.

Sireci (2013) sugiere una simplificación del modelo de validación basado en argumentos la cual propone especificar claramente los propósitos previstos (o afirmaciones) de las pruebas y que también se identifique todo uso indebido. En esta perspectiva, para construir un argumento de validez se requiere que las evidencias de las cinco fuentes de validez de la AERA et al. (2014) se combinen con las afirmaciones propuestas y las áreas de uso restringido **(Sireci, 2015)**. Mediante la recopilación de evidencias para la evaluación de cada finalidad y uso restringido de la prueba, los ingredientes para el argumento de validez se facilitan, y la evaluación y síntesis de las diversas fuentes de evidencia puede comenzar. La integración de las cinco fuentes de evidencia de validez de la AERA et al. (2014) clarifica cuáles son las evidencias que se necesitan y estandariza las características del argumento de la validez.

Mirando hacia el futuro de los conceptos de validez y validación

En este trabajo se han definido los conceptos de validez y validación para pruebas educativas y psicológicas, se ha descrito parte de la historia de la teoría de la validez y se ha explicado el proceso de validación del uso de las pruebas para un propósito determinado. La sociedad está cada vez más familiarizada con instrumentos de medida ya que se utilizan con mayor frecuencia en diversos contextos. Por esta razón, consideramos que las pruebas e instrumentos de medida serán objeto de un mayor escrutinio y que aumentará la necesidad de más exhaustivos estudios y argumentos de validez.

Consideramos que la validación puede mejorar centrándose en las consecuencias de las pruebas desde las primeras etapas de construcción, en lugar de esperar hasta después de que se administren y se utilicen para medir su impacto. Es decir, no basta con medir los impactos de las pruebas y las interpretaciones de sus puntuaciones después de que se hayan utilizado. Al contrario, se debe evaluar los diferentes sistemas de valores que subyacen las decisiones a partir de la evaluación, el constructo y el uso previsto de las puntuaciones. Esperamos que los responsables del desarrollo de las pruebas presten más aten-

ción a la hora de articular el rol adecuado de las mismas en los contextos en los que se utilizan, y que se evalúen de forma más explícita las consecuencias.

En el siglo XXI es evidente que la validez no es una propiedad de una prueba, sino que se refiere a la legitimidad del uso de una prueba para un objetivo específico **(AERA et al., 2014)**. Esperamos haber comunicado lo que los encargados del desarrollo y administración de programas de evaluación necesitan hacer, y documentar, para que los consumidores y usuarios de pruebas puedan tomar decisiones bien informadas sobre el uso apropiado de una prueba en un contexto determinado. Si existen argumentos de validez para los usos propuestos de las pruebas, las evaluaciones serán más eficaces en beneficio de la educación y la sociedad. Consideramos que la comunidad iberoamericana puede beneficiarse de las propuestas articuladas en este trabajo, ya que proporcionan una orientación sobre cómo proceder e iniciar un proceso de validación de las puntuaciones de pruebas. Este trabajo está altamente influenciado por los Estándares, que son el producto del arduo trabajo de organizaciones mundiales que se toman muy en serio el uso de las pruebas y sus consecuencias. Esperamos que, al escribir este trabajo en español sobre la teoría contemporánea de la validez, motivemos a la comunidad de hispanohablantes especialistas en medición y usuarios de pruebas a participar más plenamente en conversaciones de validez y prácticas de validación en diferentes campos de especialización.

Conclusión

Este artículo de reflexión proporcionó una reseña historia sobre la teoría de validez y destacó recientes avances que constituyen el estatus contemporáneo y evolutivo de esta teoría en el siglo XXI. Las recomendaciones propuestas en este trabajo se originan de la ardua contribución de la comunidad científica y se consideran como la postura consensual sobre los temas de validez y validación. Las líneas directrices esbozadas y detalladas son útiles para la documentación de evidencia de validez para respaldar la interpretación de las puntuaciones y usos de pruebas educativas y psicológicas. Consideramos que en el campo de educación y de psicología existe la necesidad de actualizar los estudios de validez de pruebas, para reflejar los avances actuales. Esta transición ya está en pie y especulamos que las investigaciones de validez de instrumentos de medida adoptaran con más frecuencia una perspectiva moderna de validación. Las cinco fuentes de evidencias de validez nos incitan a pensar en términos de evidencias de contenido, de constructo, de criterio y nos invita a alejarnos de la visión histórica y tripartita (validez de contenido, validez de constructo y validez de criterio).

Además de los Estándares, el marco conceptual de validación basado en la teoría de la argumentación **(Bachman y Palmer, 2010; Chapelle, 2020; Kane, 2013)** también refleja una perspectiva de validación del siglo XXI, aunque se requiere cierto manejo adecuado para su implementación. La revisión de la literatura de la teoría de la validez tanto en los Estándares como en la validación basada en argumentos recomienda la investigación de las consecuencias asociadas al uso de las pruebas. Este aspecto es particularmente importante y debe considerarse cuando las pruebas tienen un alto grado de consecuencias en la sociedad y en individuos. A medida que el alineamiento entre la teoría de validez actual y la operacionalización de validación aumente y sea más coherente, los campos de educación y de psicología podrán avanzar aún más la teoría ya que el contexto del uso de las pruebas puede informar como avanzar la teoría.

Algo cierto e inminente acerca de un artículo o libro sobre la teoría de la validez y los procesos de validación es que eventualmente terminan siendo obsoletos. Como lo han demostrado los Estándares, la teoría de la validez sigue evolucionando continuamente y es difícil predecir la trayectoria de la evolución de la teoría de la validez, sin embargo, "es probable que los cambios en los procesos de validación de pruebas sean dinámicos y paralelos a la continua madurez de nuestra ciencia" **(Geisinger, 1992, p. 219)**.

Referencias

Aguirre Forero, A. (2014). Validez del inventario de prácticas de crianza (CPC-1 versión padres) en padres madres y cuidadores de la ciudad de Bogotá. Revista Iberoamericana De Psicología, 7(1), 79-90. https://doi.org/10.33881/2027-1786.rip.7107

American Educational Research Association, Committee on Test Standards. (1955). Technical recommendations for achievement tests. American Educational Research Association.

American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (1985). Standards for educational and psychological testing. American Psychological Association.

American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (1999). Standards for educational and psychological testing. American Educational Research Association.

American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (1999). Normes de pratique du testing en psychologie et en éducation (G. Sarrazin, Trans.). Institut de recherches psychologiques.

American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

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

American Psychological Association, Committee on Test Standards. (1952). Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal. American Psychologist, 7, 461-465. https://doi.org/10.1037/h0056631

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. Psychological Bulletin, 51(2, Pt.2), 1–38. https://doi.org/10.1037/h0053479

American Psychological Association. (1966). Standards for educational and psychological tests and manuals. American Psychological Association.

American Psychological Association, American Educational Research Association y National Council on Measurement in Education. (1974). Standards for educational and psychological tests. American Psychological Association.

Anderson, D., Irvin, S., Alonzo, J., y Tindal, G. A. (2015). Gauging item alignment through online systems while controlling for rater effects: Online alignment designs and rater effects. Educational Measurement: Issues and Practice, 34(1), 22–33. https://doi.org/10.1111/emip.12038

Bachman, L. F., y Palmer, A. S. (2010). Language assessment in practice: Developing language assessments and justifying their use in the real world. Oxford University Press.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. Language Testing, 30(4), 441–465. https://doi.org/10.1177/0265532212473244

RIP 14 (1) pág. 11 -22
Revista Iberoamericana de Psicología
ISSN-L: 2027-1786

Bermúdez Jaimes, M. (2010). Diseño, construcción y análisis psicométrico de una escala de competencia social para niños de 3 a 6 años versión padres de familia. Revista Iberoamericana De Psicología, 3(1), 49-66. https://doi.org/10.33881/2027-1786.rip.3105

Bhola, D. S., Impara, J. C., y Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. Educational Measurement: Issues and Practice, 22(3), 21-29. https://doi.org/10.1111/j.1745-3992.2003.tb00134.x

Boyer, M. (2020, October 1). Fairness in educational testing [Blog post]. Disponible en https://www.nciea.org/blog/educational-assessment/fairness-educational-testing

Campbell, D. T., y Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105. https://doi.org/10.1037/h0046016

Chapelle, C. A. (2020). Argument-based validation in testing and assessment: Vol. 184. Quantitative Applications in the Social Sciences. Sage Publications Inc.

Cizek, G. J., Kosh, A. E., y Toutkoushian, E. K. (2018). Gathering and evaluating validity evidence: The generalized assessment alignment tool. Journal of Educational Measurement, 55(4), 477–512. https://doi.org/10.1111/jedm.12189

Crocker, L. M., Miller, D., y Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. Applied Measurement in Education, 2(2), 179 194. https://doi.org/ 10.1207/s15324818ame0202_6

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), Intelligence: Measurement, theory and public policy (pp. 147-171). University of Illinois Press.

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

Díaz, J., Díaz, M., y Morales, S. (2013). Diseño, construcción y validación de un instrumento que evalúa motivación laboral en trabajadores de empresas formales de la ciudad de Bogotá. Revista Iberoamericana De Psicología, 6(1), 85-94. https://doi.org/10.33881/2027-1786.rip.6109

Flora, D. B. (2020). Your coefficient Alpha is probably wrong, but which coefficient Omega is right? A tutorial on using R to obtain better reliability estimates. Advances in Methods and Practices in Psychological Science, 3(4), 484–501. https://doi.org/10.1177/2515245920951747

Garrett, H. E. (1937). Statistics in psychology and education. Longmans, Green.

Geisinger, K. F. (1992). The metamorphosis of test validation. Educational Psychologist, 27(2), 197-222. https://doi.org/10.1207/s15326985ep2702_5

Guilford, J.P. (1946). New standards for test evaluation. Educational and Psychological Measurement, 6, 427-439. https://doi.org/10.1177/001316444600600401

Hu, L., y Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 6(1), 1–55. https://doi.org/10.1080/10705519909540118

International Test Commission (2017). International Test Commission guidelines for translating and adapting tests (2nd Edition). Disponible en http://www.intestcom.org.

Jonson, J. L., Trantham, P., y Usher-Tate, B. J. (2019). An evaluative framework for reviewing fairness standards and practices in educational tests. Educational Measurement: Issues and Practice, 38(3), 6-19. https://doi.org/10.1111/emip.12259

Kane, M.T. (1992). An argument-based approach to validity. Psychological Bulletin, 112, 527-535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th edition, pp. 17-64). American Council on Education/Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50, 1-73. https://doi.org/10.1111/jedm.12000

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. Language Testing, 27(2), 183–189. https://doi.org/10.1177/0265532209349468

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. Educational Measurement: Issues and Practice, 23(4), 6–15. https://doi.org/10.1111/j.1745-3992.2004.tb00164.x

Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., y Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: Una guía práctica, revisada y actualizada. Anales de Psicología, 30(3), 1151–1169. https://doi.org/10.6018/analesps.30.3.199361

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3, 635-694 (Monograph Supplement 9).

Lovasz, N., y Slaney, K. L. (2013). What makes a hypothetical construct "hypothetical"? Tracing the origins and uses of the 'hypothetical construct' concept in psychological science. New Ideas in Psychology, 31(1), 22-31. https://doi.org/10.1016/j.newideapsych.2011.02.005

Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. Disponible en http://www.jattjournal.com/index.php/atp/article/view/45254/36645

Martone, A., y Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and instruction, Review of Educational Research 79(4), 1332-1361. https://doi.org/10.3102/0034654309341375

Messick, S. (1989). Validity. En R. Linn (Ed.), Educational measurement, (3rd ed., pp. 13-100). American Council on Education.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. W. Lissitz (Ed.), The concept of validity: Revisions, new directions, and applications (pp. 83-108). Information Age Publishing.

Mislevy, R. J. (2018). Sociocognitive foundations of educational measurement. Routledge, Taylor y Francis Group.

Mislevy, R. J., Steinberg, L. S., y Almond, R. G. (2003). Focus Article: On the Structure of Educational Assessments. Measurement: Interdisciplinary Research & Perspective, 1(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Padilla, J. y Benítez, I. (2014). Validity evidence based on response processes. Psicothema, 26(1), 136-144. https://doi.org/10.7334/psicothema2013.259

Plake, B. S., y Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, and NCME Standards for Educational and Psychological Testing? Educational Measurement: Issues and Practices, 33(4), 4-11. https://doi.org/10.1111/emip.12045

Randall, J., y Garcia, A. (2016). The history of testing special populations. In C. Wells, y M. F. Bond (Eds.), Educational measurement: From foundations to future (pp. 373-394). Guilford Press.

Riconscente, M. M., Mislevy, R. J., y Corrigan, S. (2015). Evidence-centered design. In S. Lane, M. R. Raymond y T. M. Haladyna (Eds.), Handbook of test development (2nd edition, pp. 40-63). Routledge.

Riveros Munévar, F., Bohórquez Borda, D., López Castillo, S., y Sepúlveda Rodríguez, E. (2016). Diseño y validación de un instrumento para medir las actitudes frente a la labor profesional del psicólogo. Revista Iberoamericana De Psicología, 8(2), 55 - 65. https://doi.org/10.33881/2027-1786.rip.8205

Ronquillo Horsten, L., Aranda Beltrán, C., y Pando Moreno, M. (2013). Validación de un instrumento de evaluación del desempeño en el trabajo. Revista Iberoamericana De Psicología, 6(1), 25-32. https://doi.org/10.33881/2027-1786.rip.6103

Rulon, P. J. (1946). On the validity of educational tests. Harvard Educational Review, 16, 290-296.

Russell, M., y Moncaleano, S. (2020). Examining the impact of a consensus approach to content alignment studies. Practical Assessment, Research, and Evaluation, 25(1), Article 4. Disponible en https://scholarworks.umass.edu/pare/vol25/iss1/4/

Sakakibara, K., Shimazu, A., Toyama, H., y Schaufeli, W. B. (2020). Validation of the Japanese version of the burnout assessment tool. Frontiers in Psychology, 11, 1819. https://doi.org/10.3389/fpsyg.2020.01819

Sireci, S. G. (1998). Gathering and analyzing content validity data. Educational Assessment, 5(4), 299-321. https://doi.org/10.1207/s15326977ea0504_2

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), The concept of validity: Revisions, new directions, and applications (pp. 19-37). Information Age Publishing.

Sireci, S. G. (2012, December). Smarter Balanced Assessment Consortium: Comprehensive validity agenda. Disponible en http://www.smarterbalanced.org/assessments/development/additional-technical-documentation/

Sireci, S. G. (2013). Agreeing on validity arguments. Journal of Educational Measurement, 50(1), 99-104. https://doi.org/10.1111/jedm.12005

Sireci, S. G. (2015). A theory of action for validation. In H. Jiao y R. Lissitz (Eds.). The next generation of testing: Common core standards, Smarter-Balanced, PARCC, and the nationwide testing movement (pp. 251-269). Information Age Publishing Inc.

Sireci, S. G. (2016). On the validity of useless tests. Assessment in Education: Principles, Policy y Practice, 23(2), 226-235. https://doi.org/10.1080/0969594X.2015.1072084

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. Educational Measurement: Issues and Practice, 39(3), 100-105. https://doi.org/10.1111/emip.12377

Sireci, S. G., Banda, E., y Wells, C. S. (2018). Promoting valid assessment of students with disabilities and English learners. In S. N. Elliott, J. R. Kettler, P. A. Beddow, y A. Kurz (Eds.), Handbook of accessible instruction and testing practices: Issues, innovations, and application (pp. 231-246). Sage.

Sireci, S. G., y Faulkner-Bond (2014). Validity evidence based on test content. Psicothema, 26(1), 100-107. https://doi.org/10.7334/psicothema2013.256

Sireci, S. G., y Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. Educational Measurement: Issues and Practice, 25(3), 27-34. https://doi.org/10.1111/j.1745-3992.2006.00065.x

Sireci, S. G., Rios, J. A., y Powers, S. (2016). Comparing test scores from tests administered in different languages. In N. Dorans y L. Cook (Eds.), Fairness in educational assessment and measurement (pp. 181-202). Routledge.

Smith, H. L., y Wright, W. W. (1928). Tests and measurements. Silver, Burdett.

Slaney, K. L., y Racine, T. P. (2013). What's in a name? Psychology's ever-evasive construct. New Ideas in Psychology, 31(1), 4-12. https://doi.org/10.1016/j.newideapsych.2011.02.003

Toulmin, S. E. (2003). The uses of argument (ed. revisada). Cambridge University Press.

Traynor, A. (2017). Does test item performance increase with test-to-standards Alignment? Educational Assessment, 22, 171-188. https://doi.org/10.1080/10627197.2017.1344092

Ventura-León, J. L., y Caycho-Rodríguez, T. (2017). El coeficiente Omega: Un método alternativo para la estimación de la confiabilidad. Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud, 15(1), 625-627. Disponible en https://www.redalyc.org/jatsRepo/773/77349627039/index.html

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. Applied Measurement in Education, 20(1), 7-25. https://doi.org/10.1080/08957340709336728

Wells, C. S. (2021). Assessing measurement invariance for applied research. Cambridge University Press.

Zumbo, B. (2014). What role does, and should, the test standards play outside of the United States of America? Educational Measurement: Issues and Practice, 33(4), 31-33. https://doi.org/10.1111/emip.12052

Zumbo, B. D., y Hubley, A. M. (Eds.) (2017). Understanding and investigating response processes in validation research. Springer Press.

RIP 14 (1) pág. 11 -22

Revista Iberoamericana de Psicología

ISSN-L: 2027-1786

Tabla 1. Resumen de la terminolo gía utilizada en las versiones de los Estándares

Trabajo	Terminología de validez
Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal (APA, 1952)	Categorías de validez: predictiva, estado, contenido, congruente
Technical recommendations for psychological tests and diagnostic techniques (APA, 1954)	Tipos de validez: constructo, concurrente, predictiva, contenido
Standards for educational and psychological tests and manuals (APA, 1966)	Tipos de validez: relacionadas con criterios, relacionadas con el constructo, relacionadas con el contenido
Standards for educational and psychological tests (APA, AERA, y NCME, 1974)	Aspectos de validez: relacionadas con criterios, relacionadas con el constructo, relacionadas con el contenido
Standards for educational and psychological testing (AERA, APA, y NCME, 1985)	Categorías de validez: relacionadas con criterios, relacionadas con el constructo, relacionadas con el contenido
Standards for educational and psychological testing (AERA, APA, y NCME, 1999)	Fuentes de pruebas de validez: contenido de la prueba, procesos de respuesta, estructura interna, relación con otras variables, consecuencias de las pruebas
Standards for educational and psychological testing (AERA, APA, y NCME, 2014/2018)	Fuentes de pruebas de validez: contenido de la prueba, procesos de respuesta, estructura interna, relación con otras variables, consecuencias de las pruebas