EVOLUCIÓN DEL CONCEPTO DE VALIDEZ EN LA MEDICIÓN EDUCATIVA

Adán Moisés García Medina[1]
Felipe Martínez Rizo[2]
Graciela Cordero Arroyo[3]
Joaquín Caso Niebla[4]

Introducción

La validez es considerada el criterio más importante para evaluar el rendimiento de una prueba o instrumento de medición (Cfr. AERA-APA-NCME, 2014). Su importancia es tan ampliamente reconocida que en varios países se encuentra como parte de sus leyes y reglamentos (Koretz, 2005). Sin embargo, cuando se inicia una investigación en no pocas ocasiones se suele tener una concepción limitada o incluso errónea de la magnitud y trascendencia de este concepto, y del papel que juega en el desarrollo de estudios del campo de las ciencias sociales y las humanidades.

En la actualidad hay una clara distinción entre validez y confiabilidad. La confiabilidad se refiere a la precisión con la que un instrumento de medición logra recabar la información sobre un constructo que pretende medir; se dice que hay confiabilidad en la medición cuando los resultados son estables entre grupos de individuos, o bien, en los mismos individuos a lo largo del tiempo. Por su parte, se considera que una investigación es válida cuando mide lo que pretende medir. Para un científico de las ciencias exactas esta definición podría parecer extraña o incluso absurda, sin embargo, en el campo educativo los constructos sobre los cuales se genera conocimiento suelen ser variables complejas (v. gr. rendimiento académico, clima escolar, prácticas docentes, escuela eficaz, supervisión escolar, involucramiento de los padres de familia en la escuela, etc.), compuestas por varias dimensiones. Es un reto importante observarlas a cabalidad, y cuando esto ocurre, se puede decir que la medición es válida.

La validez es una cualidad de la medición que ha sido mucho más difícil de definir. La acepción incluida en el párrafo anterior es limitada. El desarrollo de una teoría sobre la validación en instrumentos de medición ha sido paulatino, con algunos tropiezos y muchas discusiones

[1] Estudiante del Doctorado en Ciencias Educativas del Instituto de Investigación y Desarrollo Educativo de la UABC
[2] Universidad Autónoma de Aguascalientes
[3] Instituto de Investigación y Desarrollo Educativo de la Universidad Autónoma de Baja California
[4] Instituto de Investigación y Desarrollo Educativo de la Universidad Autónoma de Baja California

entreveradas. En este capítulo se pretende mostrar cuál ha sido la evolución del concepto de validez en el campo de la medición educativa, para ayudar a comprender la concepción prevaleciente y que el lector pueda identificar algunos principios básicos que se deben considerar en el desarrollo de investigaciones educativas.

El capítulo está organizado en cuatro apartados mismos que coinciden con las cuatro etapas de la teoría sobre la validación que proponen Newton y Shaw (2014). En el primero se describirán los inicios de la teoría de la validación, conocida como etapa de cristalización que ocurrió de 1921 a 1951; en el segundo apartado se tratará la etapa de la fragmentación de la validez (1952-1974); el tercero versa sobre la etapa de la reunificación, que transcurre de 1975 a 1999; y por último, en el cuarto apartado se presenta la etapa conocida como deconstrucción y que contempla de 2000 a la fecha.

1. La etapa de cristalización (de 1921 a 1951)

Aunque existen algunos desarrollos previos que provienen desde el siglo XIX, la primera época de la teoría de la validez identificada con la cristalización inició en 1921, cuando la *North American National Association of Directors of Educational Research* la incluyó entre los términos clave a definir dentro del movimiento de la medición educativa y psicológica (Newton & Shaw, 2014). Durante este periodo, la validez se definió de forma muy pragmática y poco en términos teóricos; se definía en función de la correlación de una prueba con un criterio que usualmente se asumía libre de error de medición.

Algunas definiciones representativas de la primera etapa son las siguientes: en 1937 Garrett señalaba que “la validez de un test es la fidelidad con la que mide lo que pretende medir” y Bingham definía la validez como “la correlación de las puntuaciones de un test con alguna otra medida objetiva de lo que el test quiere medir” (citados en Lissitz, 2009, p. 22). Por su parte, Guilford en 1946 la concebía así: "en un sentido general, una prueba es válida para cualquier cosa con la que se correlaciona" (citado en Messick, 1989, p. 18).

Esta visión tan empirista de la validez en buena medida se debió al desarrollo del análisis factorial que inventó Spearman a inicios del siglo XX donde se asumía que a partir de las variables observadas empíricamente se podía medir variables latentes o constructos sin considerar el error de medición. Lo anterior trajo consigo que en el campo de la psicología y la educación se desarrollasen métodos estadísticos para la validación de pruebas, tales como las técnicas correlaciones y de diferencias entre grupos.

En el capítulo sobre validez de la primera edición de una de las obras más influyente en el campo, *Educational Measurement*, Edward Cureton (1951, p. 623) la definía en términos que son muy ilustrativos de la primera etapa: “la validez de un test es la correlación entre las puntuaciones

observadas del test con las puntuaciones 'verdaderas' del criterio". Un puntaje "verdadero" es aquel que no tiene error de medición, sostenía.

Por *puntuaciones de criterio* se entendía "un conjunto de evaluaciones sin sesgo, pero no necesariamente con alta confiabilidad, de la calidad del desempeño al realizar una tarea definida, y que dicha evaluación se realizase a partir de una muestra representativa de observaciones sobre tareas de desempeño de cada miembro de una muestra representativa de personas de una población específica" (Cureton, 1951, p. 625). Sin embargo, como para variables complejas nunca es posible saber las puntuaciones reales o verdaderas, las correlaciones entre el puntaje de la prueba y el valor real son siempre estimaciones. Por ello es que el poder predictivo de una prueba se consideraba como la correlación entre las puntuaciones originales de una prueba y las puntuaciones de criterio.

Para Cureton (1951) la validez siempre debe considerarse en función de los propósitos que se persigan con las pruebas, los cuales tienen al menos dos aspectos a considerar: uno se refiere a la finalidad de lo que evalúa, el otro a la naturaleza del grupo al que se evalúa. Si se aplica un test de vocabulario a niños de sexto grado que han tenido en su trayecto escolar oportunidades e incentivos para aprender el significado de palabras escritas y provienen de hogares con características culturales similares, la prueba puede ser un indicador válido de la inteligencia verbal. Sin embargo, si la misma prueba se aplica a un grupo que proviene de hogares con capital cultural muy desigual, podría ser más válida como un indicador de la calidad de la instrucción en lectura que previamente recibieron los niños, en lugar de ser un indicador de inteligencia verbal. La validez de cualquier prueba, desde la visión de este autor, es el valor que tiene como indicador de las diferencias individuales de algún aspecto en particular entre los miembros de algún grupo específico.

A pesar de esta visión tan empirista, desde esos tiempos ya se reconocían elementos del proceso de validación que siguen vigentes hasta hoy, por ejemplo, el mismo Cureton señalaba que la pregunta fundamental de la validez de las pruebas es "qué tan bien ejecuta la función para la que fue empleada" (1951, p. 621). Lo que implica que la misma prueba puede ser usada con diferentes propósitos, y su validez puede ser alta para alguno, moderada para otro y baja para un tercer propósito. Por tanto, no se puede etiquetar a una prueba en general como de alta, moderada o baja validez, sino sólo en el contexto de ciertos fines.

En esta época ya era nítida la distinción entre validez y confiabilidad, aunque se reconocía que estaban relacionadas; Cureton (1951) señalaba que la segunda se refiere a la exactitud y consistencia con la que se mide cualquier cosa tomando en cuenta el grupo con el que se utilizará la información. En cambio, para que una prueba sea válida o que sirva adecuadamente a sus

propósitos debe medir “algo” con un razonable grado de confiabilidad, y además, ese “algo” debe ser muy similar a las operaciones que se están usando para medirlo. Lo que Cureton advertía es que la validez de un instrumento si bien depende de su relevancia y su confiabilidad, la eficacia de la prueba no es una función de la confiabilidad de las puntuaciones de criterio, sino del grado de consistencia entre los procesos que se desean medir y los que tienen que realizar los sujetos cuando se enfrentan a los instrumentos de medición.

2. La etapa de fragmentación de la validez (1952-1974)

Al inicio de la segunda etapa aparece lo que en ese tiempo se conoció como validez de contenido, en respuesta a lo restrictivo que les pareció a algunos críticos el hecho de que la validez únicamente se considerara como una correlación con conductas observables consideradas como criterio, que resultaba insuficiente para pruebas donde sus propios puntajes son el criterio a considerar. Ejemplo de ello son los test de rendimiento académico o trastornos de la conducta. Así, la validez incorporó la faceta *de contenido,* donde a partir de expertos se evaluaba si el test representaba y cubría de forma suficiente el constructo que pretendía medir.

A inicios de esta etapa aparece un famoso artículo de Cronbach y Meehl (1955) que establece que el constructo teórico, enmarcado explícitamente en una teoría, será el que determinará los datos requeridos para la validación y la interpretación de los resultados, dando así inicio a la validez de constructo. Asimismo, en 1954 la *American Psychological Association* (APA) publicó recomendaciones técnicas para la psicología y entre ellas incluyó 19 estándares para preservar la validez. Por su parte, en 1955 la *American Educational Research Association* (AERA) generó sus propios estándares para el área de la educación. En los inicios de la segunda época se distinguían cuatro tipos de validez: de contenido, predictiva, concurrente y de constructo.

En 1966 la APA y la AERA publicaron juntas un solo texto con una versión revisada y aumentada de los criterios anteriores, titulado *Standards for Educational and Psychological Tests.* Ahí se conforma lo que hasta hoy todavía algunos psicómetras denominan la trinidad de la validez, refiriéndose a lo que en ese entonces se le conoció como tres tipos de validez: de contenido, de constructo y de criterio (ésta última podría ser predictiva o concurrente). La *validez de criterio* es la que se consigue al comparar las puntuaciones de una prueba o predicciones hechas a partir de ellas, con una variable externa o criterio que provee una medición directa de las características de la conducta en cuestión. La *validez de contenido* evalúa qué tan bien el contenido de una prueba muestrea los tipos de situaciones o materias acerca de las cuales se harán las inferencias. Finalmente, la *validez de constructo* es evaluada al indagar las cualidades psicológicas o variables que realmente mide un instrumento de medición.

En el concepto de validez que propusieron la APA y la AERA en 1966 prevalece la idea de que la validación deberá considerar aspectos descriptivos y teóricos con requerimientos procedimentales y lógicos que respalden los argumentos de validación. Es decir, la validación de un instrumento implica la integración de distintos tipos de evidencia. Las pruebas son hechas con múltiples propósitos y es muy raro que un criterio sea el verdaderamente principal. El autor del capítulo dedicado a la validación de la segunda edición del *Educational Measurement,* señalaba que la mayoría de las mediciones serían inviables si uno necesitara diferentes pruebas para cada decisión o propósito (Cronbach, 1971).

De acuerdo con Cronbach (1971), la responsabilidad respecto a un uso válido de una prueba recae finalmente en la persona que interpreta los resultados. La investigación que se realiza sobre una prueba sólo provee la interpretación de algunos hechos o conceptos. Quien usa e interpreta los resultados tiene que combinar información que proporcionan esos estudios con otro conocimiento acerca de las personas que se evaluaron para decidir qué interpretaciones son factibles y válidas.

Cronbach distinguía dos grandes usos de las pruebas: para *tomar decisiones* acerca de la persona evaluada y para *describirla*. Las decisiones generalmente son realizadas para optimizar el desempeño de una persona o un grupo basado en un criterio establecido, por lo tanto, la decisión concierne a una validación de criterio; mientras que el componente descriptivo se basa en la validez de contenido y la validez de constructo. Sin embargo, la correspondencia entre estas categorías no es tan simple como aparenta; con cierta frecuencia los evaluadores que toman las decisiones sobre las personas, aplican pruebas en situaciones distintas a aquéllas en las que se realizaron los estudios de validación y para defender la extrapolación deben usar interpretaciones descriptivas. Y por otro lado, cuando una prueba se ha usado de manera descriptiva, estudios relacionados con el desempeño podrían dar luz sobre dichas interpretaciones.

Cronbach (1971) consideraba que había cinco tipos de validación, que se diferenciaban por el uso que se hacía de la prueba. Cuando el foco estaba puesto en la descripción de las interpretaciones, los tipos de validación eran: de contenido, importancia educativa y de constructo. Cuando la prueba se usaba para tomar decisiones sobre las personas a partir de un criterio, los tipos de validación eran: para la selección y para la colocación o ubicación en cierto nivel.

Los estudios de validez, según el tipo de validación en que se enfoquen, tendrían preguntas claves a responder y a partir de las respuestas se podrían hacer cierto tipo de usos. Así, para los estudios sobre *validez de contenido* la pregunta a responder es: ¿las observaciones verdaderamente representan y muestrean el universo de posibles tareas o situaciones que el

diseñador de la prueba intenta medir u observar? Este tipo de estudios serviría para decidir si las tareas o situaciones se ajustan a las categorías de contenido establecidas en las especificaciones de la prueba, así como para evaluar el proceso de la selección de contenidos.

En los estudios que se enfocan en el tipo de validación denominada *importancia educativa*, las cuestiones a responder son: ¿la prueba mide un logro educativo importante? ¿la batería de mediciones dejó de incluir algún logro educativo importante? La respuesta a estas preguntas serviría para comparar el tipo de tareas que se incluyen en la prueba con los objetivos educativos establecidos.

En las investigaciones sobre *validez de constructo*, la pregunta a responder es ¿la prueba mide el atributo que dice medir? O más específicamente, ¿la descripción de la persona en términos del constructo a medir, que se relaciona con información de esa misma persona en otras situaciones, es realmente cierta? Las respuestas permitirían: a) seleccionar las hipótesis a probar; b) integrar hallazgos para decidir si las diferencias entre las personas con altos y bajos puntajes en la prueba son consistentes con las interpretaciones que se proponen; y c) sugerir interpretaciones alternas de los datos.

En los estudios de *validez para la selección* la gran cuestión a responder es ¿las personas seleccionadas por la prueba se desempeñan mejor que aquéllas que fueron descartadas? Las respuestas servirán para: a) decidir si los criterios realmente representan los logros académicos deseados, incluyendo aquéllos logros académicos que se consiguen en un mayor periodo de tiempo; y b) decidir si una nueva situación es bastante parecida a la situación en la que se validaron los resultados que se generalizarán.

Finalmente, en las investigaciones de *validez para la colocación o ubicación en un nivel* la gran interrogante a responder es ¿el desempeño de los estudiantes mejora cuando son asignados a los tratamientos (o intervenciones educativas) de acuerdo a las puntuaciones que obtuvieron en las pruebas? Este tipo de estudios servirá para hacer los mismos juicios que en los estudios de validez para la selección.

Durante la segunda etapa, los procedimientos para examinar interpretaciones de validez se clasificaban en tres categorías: correlacionales, experimentales y lógicos (Cronbach, 1971).

Los estudios correlacionales determinan cómo difieren las personas con altos o bajos puntajes en una prueba. Así, en lo que se refiere a la convergencia de los indicadores, una persona con un alto puntaje en un test debería puntuar alto en otros indicadores del mismo constructo, o bien, en otras pruebas que midan el mismo constructo. Se advierte que en los estudios de convergencia, ninguno de los indicadores es tomado como criterio o estándar, por tanto estos trabajos no aportan a la validez de criterio sino a la de constructo. Para la validez de constructo

no basta demostrar convergencia de los indicadores, sino que se tendría que aportar evidencia de discriminación, es decir, que cierto constructo se puede distinguir de otros y eso requiere que los indicadores de un constructo tengan baja correlación con las mediciones del otro.

Los estudios experimentales buscan modificar el rendimiento de una persona en una prueba por medio de algún procedimiento controlado, para distinguir si ese elemento altera lo que se quiere medir en función de los usos que se darán a los resultados obtenidos mediante el instrumento.

Los análisis lógicos del contenido de una prueba o de las reglas de calificación pueden revelar influencias preocupantes en la puntuación. Un ejemplo simple es que ciertas medidas de logro son inválidas porque tienen un techo bajo, por ejemplo los alumnos que en una prueba previa obtienen altos puntajes, sólo pueden ganar unos pocos puntos en la prueba posterior.

No obstante que los *Standards for Educational and Psychological Tests* de 1996 y 1974 promovieron una visión trinitaria de la validez, muchos autores y usuarios interpretaron erróneamente que los tests podían validarse a través de cualquiera de los procedimientos y que era suficiente utilizar sólo uno de ellos, lo que derivó en que durante esta época se inventaran una enorme variedad de adjetivos o variantes de la validez, distintos a los propuestos por los *Standards*. En una presentación de Newton (2013), recupera de esa época casi un centenar de "tipos" de validez, donde aparecen por ejemplo, validez factorial, validez *in-situ,* validez inferencial, validez del proceso de calificación, validez funcional, validez interpretativa, etc.

3. La etapa de reunificación (1975-1999)

Ante la enorme red de términos relacionados con la validez utilizados en los años sesenta y setenta, que a su vez causaron confusión entre los diseñadores de instrumentos de medición y estudiosos del campo, teóricos como Messick (1989, 1998), Cronbach (1988) y Embretson (1983), desarrollaron argumentos para señalar que la validez de constructo es la fundamental y que el resto de acepciones forman parte de ella. La conclusión de Cronbach (1988) lo ilustra muy bien cuando sostiene: toda validación es una sola, refiriéndose a la de constructo. Por su parte, la visión de Samuel Messick, uno de los teóricos fundamentales de esa época y quien redactó el capítulo sobre validez de la tercera edición del *Educational Measurement*, es coincidente cuando afirma "la validez es un concepto unitario que siempre refiere al grado en que la evidencia empírica y el fundamento teórico apoyan lo adecuado de las interpretaciones y acciones realizadas a partir de las puntuaciones de un instrumento" (Messick, 1989, p. 13).

Para Messick (1989) aunque hay muchas maneras de acumular evidencia que apoye una inferencia en particular, esas maneras son esencialmente los métodos de la ciencia, por tanto la validación es investigación científica. Las inferencias son hipótesis y la validación de dichas inferencias equivale a probar o refutar hipótesis; sin embargo, la comprobación de hipótesis no

ocurre en forma aislada o en el vacío sino que cuando se realiza también se pone a prueba la teoría en la que se enmarcan, es decir, las fuentes, significados e importancia de las hipótesis que se generaron a partir de las puntuaciones, a su vez se derivan de teorías a partir de las cuales se interpreta el significado de los puntajes. Por tanto la validación de instrumentos debería contemplar la experimentación, la estadística y el significado filosófico mediante los cuales se evalúan las hipótesis y las teorías científicas.

En esta época Messick (1998) advertía que la validez, como la ciencia en general, no trata de explicar eventos o conductas aislados pues los fenómenos suelen estar determinados por múltiples factores. Por ello hay que considerar la consistencia de las conductas o de las respuestas a ciertos reactivos en conjunto. Advertía también que el puntaje obtenido mediante algún instrumento de medición debe ser considerado en su contexto, es decir, el contexto histórico de las personas evaluadas. Por eso la validación juega un papel fundamental, porque a través de ella se tendría que investigar de manera recurrente las interpretaciones y usos que se hagan de los resultados de un instrumento, que permitan valorar el grado en que el contexto está contaminando los resultados (Messick, 1998). En otras palabras, la validación tendría que indagar si las mediciones tienen las mismas propiedades y patrones de correlación en diferentes grupos de poblaciones y bajo diferentes condiciones.

En la evaluación educativa por lo regular se miden constructos, por lo que las puntuaciones de las pruebas son signos y muestras de dichas variables latentes. Es decir, en el campo de la medición, las conductas son consideradas como indicadores de rasgos o procesos subyacentes. Para Messick (1989) la evidencia significa tanto datos y hechos como la base lógica o argumentación que fundamenta las inferencias realizadas a partir de las puntuaciones de una prueba.

Ante la comprensión errónea de asumir que una sola de las tres categorías de validez, e incluso cualquiera de ellas, es suficiente para validar el uso de una prueba en particular, Messick (1989) propuso una forma de distinguir y combinar evidencias de validez que prevenga de una indebida dependencia de las formas en las que éstas se seleccionan, resaltando el importante papel, aunque secundario, que juegan las evidencias específicas de contenido y de criterio para apoyar la validez de constructo.

Partiendo de que la validez es un concepto unitario, Messick propuso un marco en el que se distinguían dos facetas interconectadas. Una es la justificación de la prueba, que se basa en la evaluación de evidencias o consecuencias. La otra es la función o resultado de la prueba, que es la interpretación de sus puntajes o los usos que se hacen de ellos. Al cruzar la faceta de

justificación con la de resultados, se obtienen cuatro categorías (ver Tabla 1) que se desarrollan a continuación:

Tabla 1. Facetas de la validez

RESULTADOS / JUSTIFICACIÓN	**Interpretación de la prueba**	**Uso de la prueba**
Soportada en las evidencias	Validez de constructo	Validez de constructo + Relevancia/Utilidad
Soportada en las consecuencias	Implicaciones importantes	Consecuencias sociales

Fuente: Messick (1989).

- *Justificación soportada en evidencias sobre la interpretación de pruebas*

  Para Messick (1989), la justificación basada en evidencias sobre la interpretación de los puntajes de una prueba es la validez de constructo. Por tanto la validez de constructo hace referencia, en esencia, a las evidencias y teoría que apoyan las interpretaciones del puntaje de una prueba o explicación de conceptos, los cuales toman en cuenta el desempeño en la prueba y su relación con otras variables.

  Al evaluar la manera en que se mide un constructo se deben tomar en cuenta dos consideraciones muy estrechamente relacionadas. La primera es que las pruebas son imprecisas o falibles en virtud de los errores aleatorios de medición y también porque son inevitablemente imperfectas para evaluar un constructo que suele ser una variable compleja. Las pruebas son mediciones imperfectas de un constructo porque pueden dejar fuera aspectos de este que debieron haberse incluido, porque incluyen aspectos distintos al constructo, o bien por ambas razones. La segunda es que en la validación de constructo se requieren dos tipos de evidencias, una para evaluar el grado en que ocurren las implicaciones del constructo, con la evidencia empírica obtenida de los puntajes de la prueba; otra para argumentar que esas relaciones no son atribuibles a otros constructos.

  Las evidencias del contenido aportan a la validez de constructo. En el enfoque tradicional de contenido, los ítems se incluyen en una prueba tomando como base solamente la especificación del dominio. Desde un enfoque estrictamente empírico, los reactivos se incluyen en un instrumento únicamente a partir de datos; es decir, se seleccionan aquellos ítems a partir de criterios internos como la homogeneidad de los reactivos o las cargas factoriales, o bien, mediante criterios externos como la correlación de los reactivos con alguna medición considerada criterio y la discriminación de los grupos de ítems nombrados criterio (Messick, 1989).

  Desde un enfoque sustantivo los reactivos se incluyen en un primer momento a partir de su relevancia para medir cierto dominio, pero luego son seleccionados para la prueba a partir de la

consistencia de las respuestas empíricas. El componente sustantivo de la validez de constructo es la habilidad para que la teoría que lo respalda apoye los contenidos que se exploran en una prueba.

En el marco de la validez de constructo, los modelos de las puntuaciones de las pruebas deberían ser racionalmente consistentes con la ya conocida estructura de relaciones del constructo y que está a la base de las manifestaciones de comportamiento de los sujetos evaluados. A eso Loevinger (citado en Messick, 1989, p. 43) le llamó *fidelidad estructural.* Es decir que las relaciones entre los reactivos de una prueba en su interior y con otras variables, son semejantes a las relaciones que el constructo que se quiere medir establece con esas variables.

Un aspecto de la validez de constructo es lo que Messick llama *validez nomológica,* que refiere a que la teoría del constructo a medir provee una base lógica para establecer relaciones empíricas entre los puntajes de la prueba y las mediciones de otros constructos.

Otro elemento de este tipo de validez, y que es muy parecido a la validez nomológica, es lo que Embretson (Citado en Messick, 1989, p. 48) llamó *espacio nomotético* y el cual básicamente hace referencia a las relaciones empíricas de una prueba con mediciones de otros constructos y conductas que son consideradas criterio.

- *Justificación soportada en las consecuencias de la interpretación de pruebas*

  La justificación apoyada en las consecuencias de la interpretación de la prueba consiste en evaluar las implicaciones del constructo y sus medidas asociadas. Si se considera que los constructos son conceptos amplios, y que al interpretar los puntajes de una prueba que mide algún constructo se tiene una amplia variedad de connotaciones, en el proceso de validación conviene identificar al menos tres fuentes principales: el nombre del constructo en sí, el cual debe ser evaluado de forma matizada o armónica; el valor de las connotaciones de las teorías o redes nomológicas en las que está adherido el constructo; y las implicaciones de importancia de las ideologías sobre humanidad, sociedad y ciencia que permean en la forma en que el investigador percibe y procede en su actuar (Messick, 1989).

- *Justificación soportada en evidencias sobre el uso de pruebas*

  En sentido estricto, las evidencias para justificar el uso de una prueba también pertenecen a la validez de constructo, pero requieren ser reforzadas con otras que sean específicas sobre la relevancia y utilidad de las puntuaciones de una prueba para un propósito determinado. Messick decía que usar puntuaciones de una prueba que funcionan bien sin entender lo que significan es como usar un fármaco que funciona para aliviar un malestar sin conocer sus propiedades y reacciones (Messick, 1989).

Para un adecuado uso de un instrumento es indispensable una interpretación apropiada de los puntajes. Para ello es necesario obtener evidencia convergente y discriminante que disminuya las dos mayores amenazas en la interpretación de los resultados de una prueba: a) la *subrepresentación* del constructo, asociada a la visión tradicional de la validez de contenido, y b) la varianza irrelevante al constructo, relacionada con la validación de criterio.

La *subrepresentación* del constructo en la prueba ocurre cuando un instrumento no cubre las dimensiones o facetas más importantes de un constructo. La varianza irrelevante al constructo son todos aquellos aspectos de la prueba que pueden no estar directamente relacionados con el constructo medido, pero que pueden influir en la habilidad del estudiante para contestar una pregunta específica (v. gr. lenguaje, contexto cultural).

Los especialistas distinguen dos tipos de varianza irrelevante: la que produce mayor dificultad y la que produce más facilidad. La primera ocurre cuando un agente extraño al constructo hace la prueba más difícil pero de manera irrelevante, como ocurre cuando en una prueba de una asignatura el evaluado tiene bajo nivel de lectura y es eso lo que dificulta que se mida bien su conocimiento de la materia que se evalúa. La varianza irrelevante de facilidad ocurre si un agente extraño se incluye en el reactivo o prueba y permite que algunos individuos que no tienen la habilidad en el constructo evaluado respondan correctamente el ítem, como pasa cuando un profesor, al diseñar una prueba de opción múltiple, incluye la respuesta correcta como la opción más larga y los evaluados han aprendido a lo largo del tiempo que ese maestro en particular suele diseñar sus exámenes de esa forma.

Por evidencia convergente Messick (1998) entendía aquélla que muestra con datos empíricos la misma red de correlaciones de las facetas del constructo medido que teóricamente el constructo debería mostrar. En cambio la evidencia discriminante es la que muestra que no es un constructo externo, relacionado con el que se está midiendo, el que hace que la red de correlaciones con datos empíricos coincida con lo que dice la teoría; es decir, se evalúa si existen relaciones espurias.

La varianza irrelevante asociada al constructo contamina la interpretación de los puntajes, pero no necesariamente la predicción de la prueba respecto a un criterio, porque la medida del criterio puede estar contaminada en el mismo sentido.

- *Justificación soportada en las consecuencias del uso de pruebas*

  La justificación apoyada en las consecuencias sobre el uso de pruebas consiste en considerar las consecuencias sociales de emplear un instrumento en particular para evaluar cierto constructo, como parte integral del concepto de validez.

Juzgar la validez en el sentido de si la prueba cumple el propósito para el que fue diseñada requiere que se evalúen consecuencias deseadas o imprevistas (efectos colaterales) de la interpretación de los puntajes y sus usos, considerando a los individuos en particular, a grupos de individuos, instituciones o la sociedad en su conjunto. Los efectos deben evaluarse considerando las condiciones actuales y las que pudieran ocurrir por el empleo de ciertos instrumentos en particular (Messick, 1998).

Hasta antes de los planteamientos de Messick, la omnipresencia de la validez de constructo no consideraba explícitamente la faceta en la que se evalúan las consecuencias sociales. Messick (1989) señalaba que era irónico que los estudios de validez hayan puesto poca atención a los usos y consecuencias, porque la validez en sus inicios fue concebida en términos funcionales: qué tan bien la prueba hace la tarea para la que fue diseñada.

En la literatura surgieron una serie de debates acerca de si el proceso de validación debería de incluir las consecuencias de las mediciones o evaluaciones. Algunos autores (Popham, 1997) postulaban que los límites de la validación tendrían que circunscribirse al ámbito de acción de los diseñadores de los instrumentos de evaluación; en cambio, otros argumentaban que si bien los usos que se hacían con la información obtenida mediante los instrumentos y sus posteriores consecuencias podrían estar fuera del control de los diseñadores de los instrumentos, sí tendrían que incluirse advertencias al respecto, sobre todo tratando de proteger a individuos o instituciones frente a consecuencias adversas para ellos (Cronbach, 1988; Messick, 1998).

4. La etapa de deconstrucción (2000 a la fecha)

Hacia finales de la tercera etapa, se publicó la penúltima versión de los *Standards for Educational and Psychological Testing* (AERA-NCME-APA, 1999). Dicha obra retomó la mayoría de las aportaciones de Messick que también eran coincidentes con las que posteriormente desarrollaría Michael Kane (2001, 2013) y las cuales continúan vigentes en la última edición de los *Standards* publicada a finales de 2014.

En las últimas dos ediciones de los *Standards* (AERA-APA-NCME, 2014; AERA-NCME-APA, 1999) se señala que para realizar un proceso de validación adecuado, en el que se evalúen las interpretaciones propuestas de las puntuaciones de una prueba para ciertos propósitos, se deberían obtener evidencias de cinco fuentes, que luego se concretan en estándares, que cualquier instrumento de medición debería tratar de satisfacer: de contenido, de procesos de respuesta, estructura interna, relaciones de los puntajes del instrumento con otras variables y consecuencias de la evaluación.

La primera fuente de evidencias, *de contenido,* consiste en “analizar las relaciones entre el contenido del instrumento y el constructo que se pretende medir” (AERA-APA-NCME, 2014, p.

14). En el caso de las pruebas, su contenido se refiere a las tablas de especificaciones, las preguntas que incluye la prueba, los temas que se abordan, el formato de los reactivos, las tareas que se pide efectuar a los evaluados, así como las guías de administración y calificación. Las evidencias sobre el contenido de una prueba puede incluir análisis lógicos o datos empíricos que indiquen el grado en que el contenido del instrumento representa el constructo a medir y que éste último es relevante para las interpretaciones y usos que se quiera dar a los resultados; también pueden basarse en juicios expertos.

La segunda fuente de evidencia corresponde a los *procesos de respuesta*, es decir, "un análisis empírico y teórico de la relación que existe entre los procesos que emplean los examinados para responder a las tareas que se le solicitan en determinado instrumento y su correspondencia con el constructo medido" (AERA-NCME-APA, 1999, p. 12). Si un instrumento intenta evaluar el dominio de los niños de sexto grado en ciencias naturales, se tendría que examinar si el instrumento efectivamente evalúa ciencias y no comprensión lectora. Este tipo de evidencias por lo regular proviene del análisis de respuestas individuales, y se suelen utilizar cuestionamientos a los examinados para conocer las estrategias que emplearon para resolver la tarea (entrevistas cognitivas), registros de los borradores que elaboraron hasta que generaron el escrito con el que se les evalúa, revisiones monitoreadas electrónicamente, etc.

Las evidencias sobre los procesos de respuesta también pueden provenir de un análisis de la relación entre las partes de un instrumento y éste en su conjunto, o bien de las partes de un instrumento con otras variables. Identificar diferencias individuales importantes puede servir para hacer mejoras al formato del instrumento, o bien para identificar significados o interpretaciones distintas para ciertos subgrupos de individuos a examinar.

Cuando los instrumentos requieren de jueces para calificar u observadores para registrar el desempeño, las evidencias de validez deberán considerar estudios empíricos que indiquen el grado en que los jueces u observadores son consistentes con los criterios que se han definido para calificar a los sujetos evaluados, y que no están usando criterios que son irrelevantes para las interpretaciones que se pretenden hacer con los resultados.

La tercera fuente de evidencias es sobre *la estructura interna* y consisten en "analizar el grado en que las relaciones entre los reactivos de una prueba y sus componentes corresponden a la estructura de dimensiones del constructo medido" (AERA-APA-NCME, 2014, p. 16). Los tipos de análisis y su interpretación dependen del constructo y de la forma en que serán usados los resultados del instrumento. Así, si se requiere medir un constructo que la teoría presenta como unidimensional, será adecuado emplear evidencia de la homogeneidad de los reactivos

(confiabilidad); sin embargo, si teóricamente el constructo tiene una estructura interna compleja, los índices de confiabilidad no serían una evidencia adecuada.

La cuarta fuente de evidencias es sobre *la relación con otras variables* y corresponde al "análisis de la relación de los puntajes del instrumento con otras variables externas, como mediciones de algún criterio que la prueba quiere predecir, mediciones de otros test que miden el mismo constructo, o mediciones de otros constructos que teóricamente están relacionados con el que se pretende evaluar" (AERA-NCME-APA, 1999, p. 13). Este tipo de evidencias trata de identificar el grado en que las relaciones con las otras variables son consistentes con el constructo sobre el que se pretende hacer las interpretaciones.

La evidencia puede ser convergente o discriminante. Se considera convergente cuando se analiza la relación entre puntajes de una prueba y otras puntuaciones que pretendían evaluar el mismo constructo. Se considera discriminante cuando se analiza la relación entre las puntaciones de una prueba y otras mediciones de un constructo supuestamente distinto. Las evidencias pueden provenir de estudios correlaciones o experimentales.

Asimismo, dependiendo del tipo de interpretaciones y usos del instrumento, resulta necesario analizar las relaciones del instrumento con una variable criterio. Para ello se han distinguido dos tipos de estudio: cuando la variable criterio se mide con posterioridad al constructo sobre el que se ha diseñado el instrumento y que juega el rol de predictor se denominan estudios predictivos, y cuando se obtiene información del constructo predictor y la variable criterio al mismo tiempo se denomina como estudios concurrentes.

La quinta fuente de evidencias y más reciente que surgió en respuesta a una preocupación razonable, pero que supuso una complejidad que para algunos hizo más confuso el panorama en lugar de aclararlo, fue la noción de validez *de consecuencias*, que apareció en los *Standards* AERA-APA-NCME de 1999 y dio lugar a discusiones como las que reflejan los textos que recogen dos números de la revista *Educational Measurement: Issues and Practice*: Vol. 16, N° 2, de 1997 y Vol. 17, N° 2, de 1998.

En concordancia con lo planteado por Messick (1989) y Kane (2013), quienes señalaban que los usos y decisiones tomadas a partir de una prueba deberían incluir una evaluación de las consecuencias sociales y los efectos colaterales de esas decisiones, las últimas dos ediciones de los *Standards* distinguen entre la validez de los usos de las puntuaciones de una prueba y las consecuencias de política pública que se sostienen en dichos usos. Advierten, que la información acerca de las consecuencias de una evaluación puede influir en las decisiones sobre el uso de una prueba, y que esas consecuencias por sí mismas no debilitan la validez de las interpretaciones de la prueba.

Durante esta etapa los principales teóricos de la validez (Kane, 2006; Messick, 1989) desarrollaron el *Enfoque de Validación Basado en Argumentos* (ABAV, por sus siglas en inglés), lo que se considera una aportación metodológica importante y que constituye una de las características por las que la visión predominante de la validez ha sido aceptada, en contraposición contra sus detractores quienes se han enfocado en un desarrollo eminentemente conceptual y epistemológico de la validez. Asimismo, durante esta etapa se ha puesto énfasis en la equidad de las evaluaciones (*fairness*). Por ello, enseguida se describen estos dos tópicos en los siguientes subapartados.

4.1 Enfoque de Validación Basado en Argumentos

La idea central del ABAV es establecer explícitamente y con cierto detalle (mediante argumentos) las interpretaciones que se pretende dar a los resultados de un instrumento de medición, y los usos que se piensa hacer de ellos, y luego evaluar la plausibilidad de esos propósitos (Kane, 2013). El proceso de este enfoque metodológico es simple: primero se establecen las afirmaciones que se harán a partir de las interpretaciones y usos (Kane las concibe como argumento interpretativo) y después se evalúan dichas afirmaciones a partir de las evidencias que justifican o no su uso para un propósito determinado (conocido como argumento de validez). Los Argumentos de Interpretación y Uso (IUA, por sus siglas en inglés) que se emplean en el enfoque de validación desarrollado por Kane tienen el propósito de explicitar y especificar la manera en que se interpretarán y utilizarán los resultados de una prueba con una población de determinado contexto.

Los IUA juegan el rol que tiene la teoría en el modelo de validación de Cronbach, pero de una forma más general, que permite desde interpretaciones simples, como la expectativa de desempeño en una prueba que mide cierta asignatura, hasta interpretaciones complejas sobre una teoría consolidada. Sin embargo, lo más importante de los IUA es que exponen los usos que se le pretende dar a las puntuaciones o resultados de un instrumento de medición.

El argumento de validez provee una evaluación global de las afirmaciones que se incluyen en los IUA. Las interpretaciones y usos son válidos cuando los IUA son completos y coherentes, sus inferencias razonables, y los supuestos que están a la base de las inferencias son plausibles o apoyados por evidencia adecuada. Es decir, los IUA se evalúan en términos de su claridad, coherencia y plausibilidad.

En el ABAV no tienen sentido las afirmaciones de Anastasi y Cronbach que sostienen que casi cualquier información obtenida en el proceso de desarrollo o implementación de una prueba es relevante para su validez, o que la validación es un proceso sin fin. En este enfoque, la evidencia

necesaria para la validación es aquélla que se necesita para evaluar las inferencias y los supuestos en los IUA, ni más ni menos (Kane, 2013).

Por eso los IUA tienen que ser muy específicos, de manera que se pueda valorar el tipo de evidencias que será suficiente para su evaluación. La interpretación del puntaje en una prueba de Matemáticas en términos de la habilidad para resolver cierto tipo de problemas, por ejemplo, es limitada e involucra pocas inferencias y supuestos, por lo tanto, su validación requerirá un respaldo más bien modesto. En cambio, si la interpretación de los puntajes de la misma prueba se hace en términos de la preparación y las aptitudes matemáticas de los sujetos para ingresar a una carrera, su validación requerirá mucha mayor evidencia y de distinto tipo, porque los IUA incluirán predicciones de un desempeño futuro y los argumentos de validez tendrían que evaluar la precisión de esas predicciones.

Desde un punto de vista metodológico, Kane (2013) distingue dos grandes etapas en la construcción de un instrumento y su validación: la etapa de desarrollo y la de evaluación. En la primera, que predomina al inicio del proceso, la meta es desarrollar/adaptar un programa de medición, y el foco está en desarrollar los IUA que soporten interpretaciones y usos planeados de los puntajes; en esta etapa lo importante es *la validación de las interpretaciones propuestas*.

En la segunda etapa los IUA deberían ser cuestionados, de preferencia por un evaluador neutral; si, como suele ocurrir, la validación se realiza por el equipo que diseñó la evaluación, deberían examinarse las impugnaciones que haría un crítico escéptico e incluir investigaciones empíricas de los supuestos más cuestionables de los IUA. En otras palabras, conviene adoptar una actitud confirmatoria, pero en cierto punto también se necesitará cambiar a una más crítica.

El ABAV se basa en ocho principios: 1) Lo que se valida no es la prueba en sí misma o sus puntajes, sino la interpretación de los puntajes y el uso que se haga de ellos. 2) La validez de una interpretación o uso de los resultados depende de lo bien que la evidencia apoya las afirmaciones que se hacen. 3) Afirmaciones más “ambiciosas” requieren de mayores evidencias que las soporten que aquellas afirmaciones menos “ambiciosas”. 4) Afirmaciones más “ambiciosas” (v. gr. interpretación de constructos) suelen ser más útiles que las afirmaciones menos “ambiciosas”, pero son más difíciles de validar. 5) Las interpretaciones y los usos pueden cambiar con el tiempo, en respuesta a necesidades y comprensiones nuevas, lo que conduce a cambios en las evidencias que son necesarias para la validación. 6) La evaluación del uso de los puntajes requiere una evaluación de las consecuencias de los usos planeados, y consecuencias negativas se pueden traducir en un uso inaceptable de las puntuaciones. 7) El rechazo en el uso de un puntaje no necesariamente invalida una previa interpretación de la puntuación subyacente y 8)

La validación de la interpretación de un puntaje, no valida el uso que se haga de los resultados (Kane, 2006, 2013).

4.2 La equidad en las evaluaciones

Durante esta etapa, la dimensión de justicia y equidad en las evaluaciones (*fairness)* es uno de los elementos del proceso de validación que ha ido tomando cada vez más fuerza, de hecho en las últimas versiones de los *Standards* tanto de la AERA-APA-NCME (2014) como del ETS (2014) han definido estándares específicos relacionados con este tópico e incluso han desarrollado guías específicas para promover la justicia de las evaluaciones respecto a tres fuentes de varianza irrelevantes al constructo que se pretende medir: cognitiva, afectiva y física (ETS, 2009). La equidad en la evaluación puede considerarse como la ausencia de sesgo en la medición (AERA-APA-NCME, 2014). Cualquier barrera que impida que un individuo demuestre la habilidad o constructo que se desea medir crearía un sesgo en la interpretación de los puntajes de la prueba y los usos que se hagan de ellos, lo que atentaría contra la justicia en la evaluación.

La equidad en la evaluación es un aspecto central del proceso de validación que debería cuidarse en todas las etapas de desarrollo de una prueba, la interpretación de los resultados y los usos que se hagan de ellos, es decir, es un elemento que debería ser transversal (AERA-APA-NCME, 2014; ETS, 2014). Así, la idea central de la equidad en la evaluación consiste en identificar y eliminar la varianza irrelevante de constructo para maximizar el desempeño de cualquier examinado, de tal manera que los puntajes sean equiparables entre los distintos sustentantes de la prueba.

Vale señalar que un concepto estrechamente relacionado con el de equidad en la evaluación es el de validez cultural. En relación a las evaluaciones de aprendizajes, la cultura influye en la forma en que los alumnos interpretan y responden a las actividades planteadas en las pruebas (Solano-Flores, 2011; Solano-Flores & Nelson-Barber, 2001). La falta de consideración de las características socioculturales de las poblaciones a las que están dirigidas las pruebas puede llevar a conclusiones imprecisas e inválidas sobre su desempeño. Lo anterior hace evidente la necesidad de incorporar en todo el proceso evaluativo “las influencias culturales que moldean la manera en la cual los estudiantes interpretan y resuelven problemas…” (Solano-Flores & Nelson-Barber, 2001, p. 554). Solano-Flores y Nelson-Barber definen la validez cultural como:

> …la eficacia con la que […] la evaluación aborda la influencia socio-cultural que da forma al pensamiento del estudiante y las maneras en que los estudiantes les dan sentido a […] los reactivos y sus respuestas. Las influencias socio-culturales incluyen el conjunto de valores, creencias, experiencias, patrones de comunicación, estilos de enseñanza y

> aprendizaje, epistemologías inherentes al contexto cultural de los estudiantes y las condiciones socioeconómicas que prevalecen en sus grupos culturales (citados en Basterra, Trumbull, & Solano-Flores, 2011, p. 3).

La validez cultural implica considerar la diversidad de contextos de quienes responderán los instrumentos de evaluación como una posible fuente de varianza irrelevante al constructo que se desea medir, de tal manera que los puntajes obtenidos en una prueba no se vean sesgados por aspectos ajenos al interés de la evaluación (v. gr. contexto social, cultural, económico, etc.).

En la literatura reciente hay consenso en aceptar que la atención a la diversidad en las evaluaciones debe ser atendida desde su diseño (AERA-APA-NCME, 2014; Gipps & Stobart, 2010; Solano-Flores, 2011). Para esto, es importante considerar su *accesibilidad*, entendida como la oportunidad sin obstáculos que "debe ofrecer una prueba a todos los sustentantes para demostrar el estado que tienen con respecto al constructo que pretende medir" (AERA-APA-NCME, 2014, p. 49). Se parte del supuesto que el diseño de una prueba puede impedir el acceso a demostrar el estado de un constructo, por ejemplo, cuando el diseño tipográfico no permite que los sustentantes con alguna debilidad visual puedan leer los reactivos de una prueba que no está destinada a medir la capacidad visual.

Si bien hay varios procedimientos para identificar el sesgo en las mediciones educativas, el Funcionamiento Diferencial de Reactivos (DIF, por sus siglas en inglés) ha sido una de las herramientas más empleadas con este propósito (McNamara & Roever, 2006). El DIF ocurre cuando dos individuos con el mismo nivel de habilidad tienen diferentes probabilidades de responder correctamente a un reactivo, por el hecho de pertenecer a subgrupos distintos, tales como raza, estatus socioeconómico, etnicidad, contexto lingüístico, discapacidad, contexto cultural, entre otros (Basterra et al., 2011).

Si bien existen múltiples técnicas para la detección de DIF, todas ellas tienen un elemento común: verificar que la probabilidad de respuesta correcta/incorrecta a cada ítem en un examen, sea independiente de la pertenencia de los respondentes a grupos de clasificación definidos por variables socio-demográficas y homologados por sus niveles de habilidad medida (González-Montesinos, & Jornet, 2012).

Por su parte, hay autores que recomiendan la corroboración de sesgo en los reactivos a partir de entrevistas cognitivas (Benítez & Padilla, 2014; Castillo-Díaz & Padilla, 2013; Ercikan et al., 2010). Las entrevistas cognitivas son un método general que se puede usar para evaluar el grado en que la información que se pretende medir mediante ciertos instrumentos es coincidente con lo que los sujetos deseaban proporcionar (Willis, 2015). En particular, se usan para "estudiar la manera en que la población objetivo a la que se dirigen los instrumentos son capaces de

entender, procesar mentalmente y responder a dichos instrumentos, con especial énfasis en potenciales desajustes en el proceso" (Willis, 2005, p. 3).

5. Conclusiones

La validez en el campo de la medición educativa ha ido evolucionando de tal forma que las concepciones actuales difieren mucho de las primeras. Así, durante la etapa conocida como *Cristalización,* el concepto se había enfocado a la predicción de un criterio específico, que posteriormente se conocería como validez predictiva, aunque también durante esta época se realizaron estudios de validez concurrente; ambos acercamientos utilizando métodos correlacionales, salvo que los primeros utilizando dos momentos distintos en el tiempo para la recolección de la información y en lo que respecta a los de estudios de validación concurrente, recogiendo simultáneamente datos tanto de la medida del test como del criterio con el que se comparará.

La etapa de la *fragmentación* se sigue caracterizando, salvo las aportaciones de Cronbach, por su empirismo y escasa teorización; sin embargo, uno de los avances más importantes que se lograron durante esa época fue considerar que la validez no es una propiedad del test, sino que será en función de lo adecuadas que resulten las inferencias hechas a partir de los puntajes de las pruebas o instrumentos de medición; asimismo, se logró distinguir que la validación es un proceso y que la validez es una propiedad de las inferencias, la cual admite grados, en función de la cantidad de evidencias que las soporten.

En la etapa de la *reunificación* hubo avances muy importantes en la conceptualización de la validez. Así, la validez de constructo se consideró el concepto unificador sobre el cual se adquieren diferentes tipos de evidencias de validez: contenido, procesos de respuesta, estructura interna, relación con otras variables y, usos y consecuencias. Desde una visión general, la validez se concibe como un resumen inductivo de las evidencias existentes tanto de las consecuencias potenciales a partir de la interpretación de las puntuaciones de una prueba como de sus usos. Por lo tanto, lo que debe ser validado no es la prueba o el dispositivo que se utiliza para la observación, sino las inferencias derivadas de las puntuaciones de las pruebas u otros indicadores, como las inferencias o interpretaciones del significado de una puntuación y las implicaciones que conlleva dicha interpretación.

Desde esta etapa la validez es considerada como una cuestión de grado, no de todo o nada. Además, con el tiempo, la evidencia de validez existente se fortalece o debilita por nuevos hallazgos y, además, las proyecciones de las posibles consecuencias sociales de las evaluaciones se transforman a partir de la evidencia sobre las consecuencias reales en la actualidad y las condiciones sociales cambiantes. Entonces, inevitablemente, la validez es una

propiedad en evolución y la validación es un proceso continuo. Dado que la evidencia siempre es incompleta, la validación es fundamentalmente una cuestión de tomar en cuenta los usos actuales de una prueba así como la investigación más reciente para mejorar el entendimiento de lo que significan las puntuaciones de un instrumento de medición.

Validar una inferencia interpretativa es comprobar el grado en que múltiples tipos de evidencia son consonantes con la inferencia, mientras que inferencias alternativas están menos soportadas. Validar una inferencia de acción requiere validar no sólo el significado de cierto puntaje en un instrumento de medición sino el valor de las implicaciones y de los resultados de las acciones, especialmente evaluando la relevancia y utilidad de la puntuación en una prueba para un propósito específico, así como las consecuencias sociales de usar una puntuación para la toma de decisiones (Messick, 1989).

En la etapa conocida por la *deconstrucción* de la validez pueden destacarse varios aspectos en la evolución del concepto de validez. Uno es el cambio en el énfasis, de numerosas formas específicas validez de criterio, a un pequeño número de tipos de validez, luego a una concepción unitaria de la noción y, finalmente, a distinguir cinco tipos de evidencias de validez necesarias. Otro es el cambio en el enfoque central de validación, primero centrado en la predicción y luego en la explicación, en el sentido de que la utilidad, relevancia e importancia de la predicción no pueden ser evaluadas en ausencia de una interpretación fundamentada en datos empíricos consistentes.

Para vincular el proceso de validación al de un programa de investigación científica, Kane (2001) propuso un enfoque basado en argumentos que el diseñador genera en función de los propósitos que persigue con la prueba, plantea hipótesis de trabajo y luego obtiene evidencias que respalden o refuten los argumentos e hipótesis que se planteó.

Aunque las ideas prevalecientes en la actualidad entre los estudiosos del campo y los diseñadores de pruebas han sido las establecidas en la tercera etapa, desde inicios del siglo XXI, surgió un grupo de autores como Borsboom, Mellenbergh y van Haerden (2004) y Lissitz and Samuelsen (2007) que aunque comparten varios de los postulados básicos de la visión más extendida de validez, son críticos sobre los alcances que tiene la concepción actual de la validación y difieren en aspectos epistemológicos y ontológicos.

Para estos autores, la validez sí es una propiedad del test y la definen a partir de sus características internas (confiabilidad y contenido). Aunque reconocen la importancia de otras fuentes de evidencia, para ellos las más importantes son las relacionadas con la validación de contenido, los procesos de respuesta y la confiabilidad, señalando que las otras fuentes son ajenas al proceso de validación.

Así, mientras que para los autores de la visión predominante (Kane, 2013; Messick, 1989) opinan que el diseñador de las pruebas debería cuidar que los usos que se hagan con los resultados de los test sean acordes con los propósitos para los que fue diseñado, Borsboom, Mellenbergh y van Haerden (2004) y Lissitz and Samuelsen (2007) rechazan de manera tajante dicha postura, argumentando que esas acciones escapan del campo de responsabilidades de los diseñadores de instrumentos de medición y, por tanto, no afectan la validez de un instrumento.

En la Tabla 2 se presenta de forma resumida cómo ha ido evolucionando el concepto de validez, los principales autores de cada etapa y las principales aportaciones que surgieron en cada una a la teoría de la validación.

Tabla 2 Resumen de la evolución del concepto de validez

Etapa	Concepción de Validez	Principales autores	Aportes principales a la teoría de la validación
1. Cristalización (de 1921 a 1951)	La validez centrada en la validez predictiva. La correlación entre las puntuaciones observadas del test con las puntuaciones 'verdaderas' del criterio.	Garrett Guilford Cureton	- La validez se distinguió de la confiabilidad. - Una prueba puede tener varios propósitos y su validez puede ser alta para unos de ellos y baja otros de esos propósitos.
2. Fragmentación de la validez (1952-1974)	Desde los teóricos, la validez podría ser de tres tipos: de contenido, de constructo y de criterio (ésta última podría ser predictiva o concurrente); sin embargo, entre los diseñadores de instrumentos y autores inventaron, de forma errónea, una enorme variedad de adjetivos de la validez. Newton recuperó casi un centenar "tipos" de validez usados en esta época.	Cronbach Meehl	- Surge la validez de contenido y prevalece como una faceta importante de la validez, aunque esa época se concebía como un tipo y no como una dimensión. - Surge la validez de constructo, aunque todavía como un tipo más de validez y no con la importancia que cobraría con posterioridad. - La validación de un instrumento implica la integración de distintos tipos de evidencia. Las pruebas son hechas con múltiples propósitos y es muy raro que un criterio sea el verdaderamente principal.
3. Reunificación (1975-1999)	La validez es un concepto unitario que siempre refiere al grado en que la evidencia empírica y el fundamento teórico apoyan lo adecuado de las interpretaciones y acciones realizadas a partir de las puntuaciones de un instrumento	Messick Embretson Cronbach	- Ante la gran fragmentación de la validez, los principales teóricos argumentan que toda validez es de constructo y que el resto de dimensiones forman parte de esta primera dimensión. - La validación como un proceso de investigación científica, que tendría que ocurrir en paralelo al diseño de instrumentos y posterior a ello. - La validez no como una propiedad de los test, sino de las interpretaciones de los resultados y de sus usos.

Etapa	Concepción de Validez	Principales autores	Aportes principales a la teoría de la validación
4. Deconstrucción (2000 a la fecha)	El grado en el que la evidencia y la teoría respaldan las interpretaciones de los puntajes de una prueba y los usos que se pretende hacer de ellos. Se tendría que obtener evidencias de cinco fuentes: de contenido, de procesos de respuesta, estructura interna, relaciones de los puntajes del instrumento con otras variables y consecuencias de la evaluación	Visión predominante: Kane Messick Visión crítica: Borsboom Lissitz	- Las consecuencias y usos como dimensión de primera importancia. - La equidad en las evaluaciones como un aspecto central del proceso de diseño de los instrumentos. - Metodología de validación con un Enfoque de Validación Basado en Argumentos.

Fuente: Elaboración propia a partir de los autores y textos señalados en cada uno de los subapartados.

Para finalizar es importante destacar la importancia que ha adquirido hoy el aspecto del proceso de validación que tiene que ver con las consecuencias de una medición y una evaluación. La atención que atrae este aspecto tiene que ver con la equidad, que preocupa desde la difusión del informe Coleman, pero al parecer sólo recientemente se ha vuelto una preocupación central en los procesos de validación.

Que las evaluaciones sean equitativas significa que los instrumentos no pongan en desventaja a grupos que, por sus características sociales o personales, tengan dificultad especial para desempeñarse de manera adecuada en una prueba, si esas características son ajenas a lo que se pretende medir, lo que atenta contra la esencia de la validez.

Las consecuencias de una evaluación, a las que se refiere la quinta fuente de evidencias de validez, han ocasionado debates importantes entre los teóricos del campo. En nuestra opinión es razonable la postura de los estándares de la AERA, que representan la visión que predomina entre los especialistas: los diseñadores de un test no son responsables de todos los usos que puedan hacerse de un instrumento, pero sí les corresponde advertir de manera expresa cuales son los usos adecuados y cuáles pueden ser las consecuencias de usos inadecuados que puedan vislumbrar, lo que minimizará utilizaciones inapropiadas, que pueden deberse al desconocimiento de los aspectos técnicos relevantes por parte de algunos usuarios.

Y como la validación es un proceso continuo, las sucesivas versiones de los manuales técnicos de un instrumento deberán incorporar señalamientos relativos a nuevos usos, adecuados e inadecuados, que se vayan haciendo de los resultados de una prueba. El desarrollo de instrumentos debería considerar la obtención de evidencias de validez relativas a la utilización que se pretenda dar a los resultados y, en su caso, a las consecuencias previsibles de política

pública. La obtención de tales evidencias debería ocurrir en paralelo al diseño mismo de los instrumentos de medición, en una agenda de investigación en donde se prevean los momentos en que habrá que obtener cierto tipo de evidencias, de tal manera que la construcción de los instrumentos se vaya fortaleciendo a partir de los resultados del proceso de validación.

Si a pesar de los señalamientos de los manuales ciertos usuarios emplean los resultados para decisiones o acciones que no tengan sustento sólido en ellos, dadas las características del instrumento, el mal uso será responsabilidad exclusiva del usuario.

Refiriéndose al impacto devastador que puede tener una prueba en el caso de alumnos que pese a sus esfuerzos no consiguen obtener resultados satisfactorios, Stiggins expresaba una idea que puede aplicarse a las consecuencias de cualquier evaluación:

> *Las evaluaciones más válidas y confiables del mundo que tengan como efecto hacer que los alumnos abandonen la tarea desesperanzados no pueden ser consideradas productivas, porque hacen más daño que bien [...] En el pasado, los marcos de referencia para el control de la calidad de las evaluaciones no tomaban en cuenta su impacto en el alumno; la nueva visión de la excelencia en lo relativo a evaluación, en cambio, pone en el centro de la escena este criterio de calidad*. (2008, pp. 2–3)

Referencias

AERA-APA-NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA-APA-NCME.

AERA-NCME-APA. (1999). *Standards for Educational and Psychological Testing*. American Educational Research Association; National Council on Measurement in Education; American Psychological Association.

Basterra, M. del R., Trumbull, E., & Solano-Flores, G. (2011). *Cultural Validity in Assessment*. New York; London: Routledge.

Benítez, I., & Padilla, J.-L. (2014). Analysis of Nonequivalent Assessments Across Different Linguistic Groups Using a Mixed Methods Approach Understanding the Causes of

Differential Item Functioning by Cognitive Interviewing. *Journal of Mixed Methods Research*, *8*(1), 52–68. http://doi.org/10.1177/1558689813488245

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061–1071.

Castillo-Díaz, M., & Padilla, J.-L. (2013). How Cognitive Interviewing can Provide Validity Evidence of the Response Processes to Scale Items. *Social Indicators Research*, *114*(3), 963–975. http://doi.org/10.1007/s11205-012-0184-8

Cronbach, L. J. (1971). Test Validation. En R. Thorndike (Ed.), *Educational Measurement* (2a ed., pp. 443–507). Washington, D.C: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. En Wainer, H & Braun, H (Eds.), *Test validity* (pp. 3–17). Princeton: IEA.

Cronbach, L. J., & Meehl, P. E. (1955). Construc validity in psichologycal test. *Psychological bulletin*, *52*(4), 281–302. Recuperado a partir de http://marces.org/EDMS623/Cronbach%20LJ%20&%20Meehl%20PE%20%281955%29%20Construct%20validity%20in%20psychological%20tests.pdf

Cureton, E. E. (1951). Validity. En E. Linquist, *Educational Measurement* (pp. 621–294). Washington, D.C: American Council on Education.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of Think Aloud Protocols for Examining and Confirming Sources of Differential Item Functioning Identified by Expert Reviews. *Educational Measurement: Issues and Practice*, *29*(2), 24–35. http://doi.org/10.1111/j.1745-3992.2010.00173.x

ETS. (2009). ETS Guidelines for Fairness Review of Assessments. Educational Testing Service. Recuperado a partir de https://www.ets.org/Media/About_ETS/pdf/overview.pdf

ETS. (2014). *ETS Standards for Quality and Fairness*. Princeton, NJ: Educational Testing Service.

Gipps, C., & Stobart, G. (2010). Fairness. En *International Encyclopedia of Education* (3a ed., Vol. 4, pp. 56–60). Oxford, UK.: Elsevier.

González-Montesinos, M., & Jornet, J. M. (2012). Procedimientos para la Detección del Funcionamiento Diferencial de Reactivos (DIF). Documento preparado para el INEE.

Kane, M. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement*, *38*(4), 319–342. Recuperado a partir de http://www.jstor.org/stable/1435453

Kane, M. (2006). Validation. En R. L. Brennan (Ed.), *Educational Measurement* (4a ed., pp. 17–64). Washington, D.C: ACE-NCME.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. Recuperado a partir de http://onlinelibrary.wiley.com/doi/10.1111/jedm.12000/full

Koretz, D. (2005). *Alignment, High Stakes, and the Inflation of Test Scores* (No. 655). Los Angeles, CA: Center for the Study of Evaluation (CSE), National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Graduate School of Education & Information Studies, University of California, Los Angeles.

Lissitz, R. W. (Ed.). (2009). *The concept of validity. Revisions, New Directions and Applications*. Charlotte: Information Age Publ.

Lissitz, R. W., & Samuelsen, K. (2007). A Suggested Change in Terminology and Emphasis Regarding Validity and Education. *Educational Researcher*, *36*(8), 437–448. http://doi.org/10.3102/0013189X07311286

McNamara, & Roever. (2006). Psychometric Approaches to Fairness: Bias and DIF. *Language Learning*, *56*(S2), 81–128. http://doi.org/10.1111/j.1467-9922.2006.00381.x

Messick, S. (1989). Validity. En R. L. Linn, *Educational Measurement* (3a ed., pp. 13–103). New York: ACE-NCME.

Messick, S. (1998). Test Validity: A Matter of Consequence. *Social Indicators Research*, *45*(1-3), 35–44. http://doi.org/10.1023/A:1006964925094

Newton, P. (2013, febrero). *Does it matter what "validity" means?* Presentado en Seminar University of Oxford, Department of Education, Oxford.

Newton, P., & Shaw, S. (2014). *Validity in Educational and Psychological Assessment*. SAGE.

Popham, W. J. (1997). Consequential validity: Right Concern-Wrong Concept. *Educational measurement: Issues and practice*, *16*(2), 9–13. Recuperado a partir de http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.1997.tb00586.x/abstract

Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices. En M. del R. Basterra, E. Trumbull, & G. Solano-Flores, *Cultural validity in assessment* (pp. 3–21). New York: Routledge.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38*(5), 553–573.

Stiggins, R. (2008). *Assessment Manifesto: A call for the development of balanced assessment systems*. Portland, Oregon: ETS-ATI. Recuperado a partir de http://www.uvstorm.org/Downloads/AssessManifesto-08.pdf

Willis, G. B. (2005). *Cognitive interviewing: a tool for improving questionnaire design*. Thousand Oaks, Calif.: Sage Publications.

Willis, G. B. (2015). *Analysis of the cognitive interview in questionnaire design*. Oxford: Oxford University Press.