EVIDENCIAS SOBRE LA VALIDEZ DE CONTENIDO: AVANCES TEÓRICOS Y MÉTODOS PARA SU ESTIMACIÓN

CONTENT VALIDITY EVIDENCES: THEORETICAL ADVANCES AND ESTIMATION METHODS

Ignacio Pedrosa, Javier Suárez-Álvarez y Eduardo García-Cueto

Universidad de Oviedo

Cómo referenciar este artículo/How to reference this article:

Pedrosa, I., Suárez-Álvarez y García-Cueto, E. (2013). Evidencias sobre la Validez de Contenido: Avances Teóricos y Métodos para su Estimación [Content Validity Evidences: Theoretical Advances and Estimation Methods]. *Acción Psicológica, 10*(2), 3-18. http://dx.doi.org/10.5944/ap.10.2.11820

Resumen

La finalidad de este trabajo ha sido realizar una revisión sobre la evolución histórica de la validez de contenido, así como presentar algunos de los métodos de estudio más utilizados para su estimación. El concepto de validez de contenido ha sido objeto de un largo proceso de modificaciones desde su origen. Sin embargo, estos cambios han estado focalizados en qué el tipo de evidencias se deben presentar para su estudio y los métodos más adecuados para encontrar dichas evidencias. Sin embargo, su definición se ha mantenido estable a lo largo del tiempo. En la actualidad, la validez de contenido se considera condición necesaria (aunque no suficiente) para realizar interpretaciones de las puntuaciones en los tests. Finalmente, la combinación de métodos tanto cualitativos como cuantitativos se entiende como el procedimiento más completo a la hora de realizar un estudio de validez de contenido en profundidad. Dentro de los primeros, destaca el índice IVC como el que, además de ser el más empleado actualmente, presenta los mayores beneficios respecto a las diferentes alternativas propuestas a lo largo de los años. Respecto a los segundos, la Teoría de la Generalizabilidad se entiende como el procedimiento más exhaustivo y cercano a la estimación de la validez de contenido en sí.

Palabras Clave: Validez, Validez de contenido, Estimación, Constructo.

Abstract

The purpose of this paper has been to carry out a review of the historical evolution of one aspect of test validity - content validity - as well as to expose some of the most popular methods used for its evaluation. The concept of content validity has undergone many modifications from its inception to the present time. However, in the past these changes have focused on which pieces of evidence must be presented and the diffe-

Correspondencia: Ignacio Pedrosa, Facultad de Psicología, Universidad de Oviedo, Plaza Feijoo, s/n, Cabina 4, 33003, Oviedo, España. Email: pedrosaignacio@uniovi.es.

Recibido: 19/02/2013
Aceptado: 12/06/2013

rent statistical methods used to study it, while keeping its meaning essentially stable over time. Nowadays, evidence for content validity is considered necessary (though not the sole factor) for interpreting test results. Finally, the use of both qualitative and quantitative methods is recognized as the best procedure for performing an in-depth study of content validity. Regarding the former, this paper recommends the IVC index because, besides being currently the most used, it shows the most benefits when compared to the alternatives proposed over the years. Regarding the latter, the Generalizability Theory is understood as the most comprehensive and accurate procedure for measuring this aspect of test validity.

Keywords: Validity, Content validity, Estimation, Construct.

Introducción

Un test es un instrumento de medida a partir del cual es posible obtener una muestra de conducta sobre la que se pretenden hacer ciertas inferencias, mientras que el concepto de validez se refiere al conjunto de pruebas y datos que han de recogerse para garantizar la pertinencia de tales inferencias (Muñiz, 2000). Según la edición más reciente de los «Estándares para el uso de tests psicológicos y educacionales» (American Educational Research Association [AERA], American Psychological Association [APA], y National Council on Measurement in Education [NCME], 1999), «validez se refiere al grado en que la evidencia y la teoría apoyan las interpretaciones de las puntuaciones en los tests» (1999, p. 9). Más concretamente, «se validan las inferencias relativas al uso específico de un test, y no el propio test» (AERA, APA y NCME, 1985, p. 9). Es decir, no existen tests válidos sino que los tests son válidos para algo, siendo indispensable indicar a los usuarios potenciales del test sus limitaciones así como concretar para qué es válido exactamente.

Como sintetiza Sireci (2009), las fuentes de evidencia de validez han sufrido un proceso de «embalaje» y «desembalaje». En este sentido, parece ser que la tendencia a lo largo de la historia de la validez hasta la actualidad ha sido un «desembalaje» (hacia varios tipos de validez) y un «embalaje» (hacia una conceptualización unitaria), siendo además previsible que esta metamorfosis continúe en el futuro. En cualquier caso, no hay duda que la concepción actual es tomar la validez como única, existiendo diferentes fuentes para probar dicha validez. En este sentido, las recomendaciones técnicas de las comisiones internacionales sugieren cinco fuentes de evidencia de validez: contenido, procesos de respuesta, estructura interna, relaciones con otras variables y consecuencias de la evaluación (AERA, APA y NCME, 1999).

Aproximación al concepto de validez de contenido

Dentro de la validez de contenido en sí, los trabajos de Rulon (1946), Mosier (1947) y Gulliksen (1950a, 1950b) podrían considerarse los prolegómenos sobre los que surge el concepto acerca de este tipo de validez (Sireci, 1998a). Sin embargo, la primera aproximación a una definición operativa podría tener su origen en Cureton (1951).

Cureton presentó una novedosa definición de validez de contenido que supuso la introducción del término en la literatura sobre pruebas educativas y psicológicas (Sireci, 1998a). Su principal aportación es el reconocimiento de la existencia de una relevancia curricular o validez de contenido. En este sentido, afirma que si se pretenden validar ítems estadísticamente, se tendría que poder aceptar que el criterio de trabajo es adecuado. Para ello, los ítems «tendrían que evocar aquello que dicen estar midiendo y constituir una muestra representativa del universo de medida» (Cureton, 1951, p. 664). Una vez establecido este sustento teórico, es cuando surgen los dos criterios fundamentales para estudiar la validez de contenido: relevancia y representatividad.

El concepto de validez de contenido ha sido objeto de múltiples transformaciones desde sus orígenes. Sin embargo, estos cambios han estado más bien focalizados en otorgarle importancia como fuente de evidencia de vali-

dez que en su definición operativa, la cual ha permanecido esencialmente estable desde su origen. Así por ejemplo, Anastasi (1954) describió la validez de contenido como especialmente pertinente para la evaluación de pruebas de rendimiento. Sin embargo, no apoyaba su uso a la hora de validar tests de aptitudes o de personalidad. Cronbach y Meehl (1955), aunque diferenciaban entre validez de criterio, de contenido y de constructo, enfatizaban esta última considerándola aplicable a todos los tests. Por el contrario, Ebel (1956), resaltó la importancia de la validez de contenido hasta el punto de considerarla como la base de la validez de constructo.

Paralelamente a estas disquisiciones, la APA comenzaba a referirse al contenido de los tests en sus publicaciones sobre las recomendaciones técnicas para el diseño y uso de los tests. La tendencia histórica de la validez de contenido desde las primeras «Recomendaciones técnicas para los tests psicológicos y técnicas diagnósticas» (APA, 1952) hasta los últimos «Estándares para el uso de tests psicológicos y educacionales» (AERA, APA y NCME, 1999), ha sido el incremento de su protagonismo, convirtiéndose actualmente en una de las principales fuentes de evidencias de validez.

Respecto a su definición, Guion (1977), realiza una definición operativa basada en cinco condiciones que considera necesarias para aceptar una medida en función de su contenido:

1. El contenido del dominio debe tener sus raíces en la conducta, con un significado generalmente aceptado.
2. El contenido del dominio debe ser definido sin ambigüedad.
3. El contenido del dominio debe ser relevante para los objetivos de medida.
4. Jueces cualificados deben estar de acuerdo en que el dominio ha sido adecuadamente muestreado.
5. El contenido de las respuestas debe ser observado y evaluado de forma fiable.

Este planteamiento se aproxima a las perspectivas más actuales. Como describe Kane (2006, p. 149), las primeras dos condiciones sugieren la necesidad de un dominio bien definido. Su primera y tercera condición requiere que el dominio sea relevante para la interpretación propuesta así como para el uso de las puntuaciones en el test. Su cuarta condición alude al muestreo representativo y la última de ellas requiere tanto puntuar de forma precisa como que las puntuaciones observadas sean generalizables. A continuación se presenta una selección de publicaciones que permiten profundizar en la evolución de la conceptualización de la validez de contenido a lo largo de su historia (ver Tabla 1).

Tabla 1

Publicaciones sobre la definición de los aspectos de la validez de contenido

Representación del Dominio	Relevancia del Dominio	Definición del Dominio	Procedimientos de construcción
Mosier (1947)	Rulon (1946)	Thorndike (1949)	Loevinger (1957)
Goodenough (1949)	Thorndike (1949)	APA (1952)	Ebel (1956, 1961)
Cureton (1951)	Gulliksen (1950a)	Lennon (1956)	AERA/APA/NCME(1966)
APA (1952)	Cureton (1951)	Ebel (1956, 1961)	Nunnally (1967)
AERA/APA/NCME (1954)	AERA/APA/NCME (1954)	AERA/APA/NCME (1966)	Cronbach (1971)
Lennon (1956)	AERA/APA/NCME (1966)	Cronbach (1971)	Guion (1977, 1980)
Loevinger (1957)	Cronbach (1971)	AERA/APA/NCME (1974)	Tenopyr (1977)
AERA/APA/NCME (1966)	Messick (1975, 80, 88, 89a, b)	Guion (1977, 1980)	Fitzpatrick (1983)
Nunnally (1967)	Guion (1977, 1980)	Tenopyr (1977)	AERA/APA/NCME (1985)
Cronbach (1971)	Fitzpatrick (1983)	Fitzpatrick (1983)	
AERA/APA/NCME (1974)	AERA/APA/NCME (1985)	Messick (1975, 80, 88, 89a, b)	
Guion (1977, 1980)			
Fitzpatrick (1983)			
AERA/APA/NCME (1985)			

Fuente: Sireci (1998a, p. 102)

En la actualidad, la validez de contenido se considera condición necesaria (aunque no suficiente) para realizar interpretaciones de las puntuaciones en los tests (Kane, 2009, p. 61). Además, ésta no se refiere únicamente a los ítems del instrumento de medida, sino que también incluye las instrucciones para su administración y los criterios para su corrección y puntuación (Abad, Olea, Ponsoda y García, 2011).

Sireci (2003) indica que hay, al menos, dos fuentes principales de evidencias de validez de contenido: la definición del dominio y la representación del dominio. La definición del dominio se refiere a la definición operativa del contenido (i.e. tabla de especificaciones). El segundo elemento, la representación del dominio, abarca tanto la representatividad como la relevancia. Dentro de este segundo elemento, la representatividad indica la adecuación con que el contenido del test representa todas las facetas del dominio definido, mientras que la relevancia alude al grado en que cada ítem del test mide el dominio definido, pudiéndose detectar contenidos irrelevantes.

Métodos y aplicaciones para la estimación de la validez de contenido

Según Sireci (1998a), se pueden establecer dos planteamientos para estimar la validez de contenido: métodos basados en el juicio de expertos y la utilización de métodos estadísticos derivados de la aplicación del instrumento de medida.

Si bien resultaría excesivamente ambicioso pretender aglutinar en el presente estudio la totalidad de métodos existentes para estimar la validez de contenido, a lo largo de las siguientes páginas se trata de exponer, a modo de evolución histórica, aquellos que presentan o han tenido una mayor difusión y aplicación a nivel práctico.

Métodos basados en el juicio de expertos

Estos métodos se caracterizan por contar con un número de expertos que bien proponen los ítems o dimensiones que deben conformar el constructo de interés o evalúan los diferentes ítems en función de su relevancia y representatividad, en base a una escala tipo Likert, y emiten juicios sobre el grado de emparejamiento entre los elementos y los contenidos que han de ser evaluados (Abad, et al., 2011).

En este punto, antes de profundizar en los diferentes métodos existentes, se considera relevante destacar dos aspectos que se entienden como determinantes a la hora de evaluar la validez de contenido de un instrumento.

En primer lugar, la apropiada selección de los expertos supone una cuestión fundamental a la hora de establecer este tipo de validez. Por ello, si se pretende realizar un adecuado análisis de los elementos, resulta fundamental analizar las características y experiencia de los expertos en relación al constructo tratado. Una interesante reflexión en torno a este tema puede consultarse en Lawshe (1975).

Por otro lado, tradicionalmente, el procedimiento de evaluación por parte de los expertos ha consistido en que estos, conociendo las dimensiones que se pretende evaluar, valoren y asignen cada uno de los ítems a dichas dimensiones (Sireci, 1998b). Este tipo de instrucciones puede introducir importantes sesgos, ya que si conocen qué se pretende medir y estos constructos vienen definidos por el propio investigador, existe el riesgo de «dirigir» la valoración, pudiendo provocar un incremento artificial de las tasas de utilidad y relevancia del ítem y alterando así la información real acerca del instrumento. A pesar de la importancia de este posible sesgo, son escasos los métodos objetivos desarrollados para evitar este problema, siendo las combinaciones binarias de Thurstone (1927), uno de los más adecuados a nivel práctico. Obviamente, este método cuenta con el problema de que el número de ítems sea excesivamente elevado, derivando en un número de combinaciones excesivamente grande.

Al margen del análisis cualitativo de los expertos, resulta imprescindible que estos aporten una valoración cuantitativa a los ítems. En caso contrario, el mero hecho de que informen sobre la falta o exceso de ítems representativos del constructo o que simplemente determinen a qué dimensión corresponde cada elemento, no aporta de por sí información relevante para el proceso de validación (Sireci, 1998a). Por esta razón, es fundamental aplicar alguno de los métodos empíricos existentes para cuantificar este grado de acuerdo.

Así pues, retomando la senda de los procedimientos existentes, se ha comentado anteriormente cómo la valoración de los expertos suele realizarse en base a una escala tipo Likert. Estas escalas pueden presentar ligeras modificaciones, bien en cuanto al número de alternativas empleadas, las propuestas varían entre las cinco alternativas (Mussio y Smith, 1973) y las tres planteadas por Hambleton (1980), o bien en cuanto a la tarea en sí, solicitando valorar aspectos como la utilidad, relevancia, importancia, etc. de cada elemento (Drauden y Peterson, 1974). Al margen de estas ligeras diferencias, todas ellas presentan como objetivo fundamental decidir en qué medida el ítem se ajusta al constructo de interés.

En este sentido, los métodos propuestos han sido diversos y se han incrementado paulatinamente a lo largo de los años. Así, realizando un recorrido histórico, se puede considerar a Tucker (1961) como el precursor en este campo.

Método basado en el Análisis Factorial (Tucker, 1961)

El método planteado por Tucker se basa en el análisis factorial de las puntuaciones otorgadas por los expertos en cuanto a la relevancia de los ítems, pudiendo obtener dos factores diferenciados. El primero de ellos, puede interpretarse como una adecuación muestral de los ítems para constituir un test, al considerar el test como una muestra representativa de la va-

riable de interés. Por otra parte, el segundo permite detectar las diferencias de puntuaciones dadas en la evaluación de los expertos.

Índice de Validez de Contenido (Lawshe, 1975)

Tras un considerable número de años sin avances a nivel cuantitativo en esta materia, es Lawshe quien propone uno de los índices más conocidos de todos los desarrollados en este campo, el cual fue denominado como IVC. Lawshe, desde una orientación de la Psicología del Trabajo y las Organizaciones, planteó en su trabajo «Quantitative approach to content validity» (1975) un índice empírico para relacionar el contenido de un instrumento de selección de personal con el desempeño laboral.

Este método, conocido como Panel de Evaluación del Contenido, consiste en la evaluación individual de los ítems de un test por parte de un grupo de expertos en la materia. A continuación, mediante la Razón de Validez de Contenido (RVC, *Coefficient Validity Ratio* en inglés), se determina qué ítems del instrumento son adecuados y deben mantenerse en la versión final del mismo. En este punto, se debe asignar a cada ítem una puntuación en base a tres posibilidades: que el elemento sea esencial para evaluar el constructo, que resulte útil, pero prescindible o que se considere innecesario. Sobre esta valoración se aplica la siguiente expresión:

$$RVC = \frac{n - N/2}{N/2}$$

donde n es el número de expertos que otorgan la calificación de esencial al ítem y N, el número total de expertos que evalúan el contenido.

Finalmente, se calcula el Índice de Validez de Contenido (IVC, *Content Validity Index* en inglés) para el instrumento en su conjunto, el cual no es más que un promedio de la validez de contenido de todos los ítems seleccionados en el paso previo.

En cuanto a la interpretación de este índice, existen dos tendencias en función de que se adopte un criterio más o menos flexible. Así, por un lado, es posible interpretarlo bien a nivel de significación estadística, teniendo que ser el IVC superior a una probabilidad asociada de 0.05 (Lynn, 1986) o bien, como propone Davis (1992), interpretando directamente el índice obtenido y teniendo que ser superior a 0,80 para definir el conjunto de ítems como adecuado. Sin embargo, desde otra perspectiva menos estricta, autores como Rubio, Berg-Weber, Tebb, Lee y Rauch (2003), proponen que el grado de acuerdo esperado en torno a un ítem se ajuste al número de expertos que participan en la evaluación. Para ello, el propio Lawshe elaboró una tabla que relaciona los valores obtenidos en este índice y el número de expertos empleado. De este modo, el valor crítico de la RVC se incrementa de manera monotónica cuando se emplean entre 40 y 9 expertos (siendo los valores mínimos adecuados de.29 y.78, respectivamente) y alcanzando el grado máximo de acuerdo (.99) cuando se recurre a 7 expertos o menos.

Una interpretación similar es la aportada por Lynn (1986), quien establece el valor mínimo del índice teniendo en cuenta el número de expertos participantes y el número de expertos que consideran el ítem como relevante. En esta misma línea, otros investigadores han propuesto puntos de corte valorando, al mismo tiempo, el número de elementos evaluados, la consistencia interna de las escalas de evaluación e, incluso, las implicaciones prácticas de los instrumentos de medida (Crocker, Llabre y Miller, 1988).

Ejemplos de aplicación directa de este índice pueden consultarse en numerosos trabajos aplicados a diferentes áreas como los de Bazarganipour, Ziaei, Montazeri, Faghihzadeh y Frozanfard (2012) en el ámbito clínico, Castle (2008) en el entorno laboral o Yeun y Shin-Park (2006) a la hora de analizar la validez transcultural de un instrumento.

Índice de congruencia ítem-objetivo (Rovinelli y Hambleton, 1977)

Una aportación afín al IVC es la presentada por Rovinelli y Hambleton (1977) mediante el

índice de congruencia ítem-objetivo. Para ello, el juez debe valorar como +1 o -1 según el ítem mida o no el objetivo deseado y, aplicando sobre estos datos, la siguiente expresión:

$$I_{jk} = \frac{N}{2N - 2} (\bar{X}jk - \bar{X}j)$$

siendo N el número de objetivos, la media de los jueces para el ítem j en el objetivo k y la media para el ítem j en todos los objetivos.

A partir de aquí, debe fijarse el grado de acuerdo mínimo esperado por el investigador para seleccionar los ítems adecuados. Aplicaciones prácticas de este índice pueden consultarse en trabajos como los de García-Campayo et al. (2009) o García-Campayo et al. (2012).

Índice de congruencia (Hambleton, 1980, 1984)

De forma progresiva siguen apareciendo nuevos métodos, surgiendo, por ejemplo, un nuevo índice propuesto por Hambleton (1980) basado, en este caso, en una perspectiva centrada en los test referidos al criterio. A partir de este tipo de tests, planteó el denominado índice de congruencia ítem-objetivo, según el cual compara el grado en que un ítem evalúa el constructo esperado en relación al resto de dimensiones que componen el instrumento.

Más adelante, el propio Hambleton (1984), propuso una variación de su método con el objetivo tanto de facilitar la labor de los expertos como de poder obtener éste índice independientemente del número de alternativas empleadas para evaluar los ítems. Así, además de la relación de cada ítem respecto al constructo, es posible obtener un índice de congruencia que describa el ajuste de cada ítem respecto al instrumento total teniendo en cuenta las valoraciones de la totalidad de expertos.

V de Aiken (Aiken, 1980)

De manera paralela, Aiken (1980), elaboró un índice similar al establecido por Hambleton (1980). Dicho índice permite evaluar la relevancia de cada ítem respecto a su constructo; pero teniendo en cuenta, en este caso, no sólo el número de categorías ofrecidas a los jueces, sino también el número de expertos participantes. Sobre estos datos, se establece el grado de acuerdo basado en la distribución normal y obteniendo, a partir de ella, una probabilidad asociada a cada ítem (para profundizar en el cálculo de este índice, consultar Merino y Livia, 2009). Una aplicación práctica de este índice a una escala destinada a valorar el desempeño laboral puede consultarse en Distefano, Pryer y Erffmeyer (1983).

Por supuesto, en ambos casos, al igual que ocurre en los diferentes métodos que se presentarán más adelante, es posible obtener una valoración global del instrumento diseñado.

Escalamiento multidimensional y análisis de clusters (Sireci y Geisienger, 1992)

Una década más tarde, estos autores establecen un método en una línea diferente. Así, pretenden valorar la tasa de similaridad de los ítems basándose en el escalamiento multidimensional y el análisis de clusters. Este procedimiento supone, además de un cambio en la perspectiva de análisis de los datos aportados por los expertos, una solución al problema previamente señalado sobre el sesgo introducido en la investigación cuando los expertos conocen las especificaciones del contenido que se pretende valorar.

El planteamiento consiste en presentar el conjunto de ítems a los expertos para que sean estos quienes los asocien en base a su similaridad. La lógica subyacente es aquellos ítems similares serán agrupados conjuntamente formando un mismo *cluster* y se encontrarán, a su vez, muy próximos entre sí a la hora de realizar el escalamiento multidimensional. La combinación de ambos resultados permite analizar así la convergencia/divergencia de los constructos obtenidos.

En un estudio de estos mismos autores (Sireci y Geisienger, 1995), puede verse la aplica-

ción del método a dos cuestionarios para la evaluación de habilidades cognitivas.

Poco después, Deville (1996) amplió este método teniendo en cuenta tanto la relevancia otorgada a cada ítem como las respuestas de los participantes a cada elemento y aplicando el escalamiento multidimensional sobre estos datos. Con esta propuesta, Deville, va un poco más allá relacionando tanto la validez de contenido como de constructo.

Método de Capacidades Mínimas (Levine, Maye, Ulm y Gordon, 1997)

Al igual que ocurría con el método propuesto por Hambleton (1980), a finales de siglo, Levine et al. (1997) formulan un nuevo método basado en los test referidos al criterio y, concretamente, en la selección de personal. Este método, conocido como Capacidades Mínimas (*Minimum qualifications, MQs*, en inglés), presenta como característica la focalización en el nivel de capacidad o habilidad mínima necesaria para tener éxito en un determinado criterio.

Para ello, establecen, en primer lugar, un perfil de las características que cada trabajador debe poseer en relación a su rol laboral. Posteriormente, un panel de expertos define, mediante el método de Angoff (1971), el nivel de habilidad mínimo que el empleado debe poseer para cumplir con el perfil propuesto. Finalmente, estos expertos evalúan, por un lado, cada tarea en cuanto a la dificultad de alcanzar cada una de las capacidades mínimas y, por otra parte, el nivel de cada aspirante en relación a las tareas propuestas. De este modo, se selecciona a quienes cumplen un nivel mínimo en las tareas que se entiende definen el constructo (criterio) en que deberán tener éxito.

A pesar de que el planteamiento inicial de este método era eminentemente laboral, su metodología permite que sea aplicable a otros contextos de evaluación. Una muestra de ello, es su aplicación al ámbito educativo propuesta por Buster, Roth y Bobko (2005) quienes, introduciendo ciertas modificaciones, ejemplifican la adecuación de este método a un contexto diferente.

Rango Interpercentil Ajustado a la Simetría (Fitch, et al., 2001)

Para la aplicación de este método (conocido como *IPRAS* en inglés), los expertos deben valorar, en una escala tipo Likert de 9 puntos, la adecuación y relevancia de los distintos ítems. Posteriormente, para mantener el ítem en el instrumento final éste debe, en primer lugar, presentar una mediana superior a 7 y, a continuación, existir un acuerdo entre los distintos expertos acerca del ítem. En este segundo punto es donde se calcula el rango interpercentil (*IPR*, en inglés) como medida de dispersión (idealmente entre el 30 y el 70%).

Finalmente, este rango calculado (IPR) debe ser comparado con el IPRAS, seleccionando el ítem si el IPRAS asume un valor superior al IPR. En el estudio de Kröger et al. (2007), puede analizarse su aplicación a una escala destinada a evaluar el daño cognitivo en personas mayores.

Coeficiente de Validez de Contenido (Hernández-Nieto, 2002)

Otra propuesta es el Coeficiente de Validez de Contenido (CVC; Hernández-Nieto, 2002). Al igual que los coeficientes clásicos ya expuestos, éste permite valorar el grado de acuerdo de los expertos (el autor recomienda la participación de entre tres y cinco expertos) respecto a cada uno de los diferentes ítems y al instrumento en general. Para ello, tras la aplicación de una escala tipo Likert de cinco alternativas, se calcula la media obtenida en cada uno de los ítems y, en base a esta, se calcula el CVC para cada elemento.

Así,

$$CVC_i = \frac{M_x}{V_{máx}}$$

donde M_x representa la media del elemento en la puntuación dada por los expertos y $V_{máx}$ la puntuación máxima que el ítem podría alcanzar. Por otro lado, debe calcularse el error asignado a cada ítem (Pe_i), de este modo se reduce el posible sesgo introducido por alguno de los jueces, obtenido mediante

$$Pe_i = \left(\frac{1}{j}\right)^j$$

siendo j el número de expertos participantes. Finalmente, el CVC se calcularía aplicando $CVC = CVC_i - Pe_i$.

Respecto a su interpretación, Hernández-Nieto (2002) recomienda mantener únicamente aquellos ítems con un CVC superior a 0.80, aunque algunos criterios menos estrictos establecen valores superiores a 0.70 (Balbinotti, 2004). El trabajo de Balbinotti, Benetti y Terra (2007), presenta la aplicación de este método a la hora de traducir y adaptar una escala centrada en el contexto financiero.

Índice de Validez Factorial (Rubio et al., 2003)

Otro de los métodos relativamente reciente es el desarrollado por Rubio et al. (2003). Esté método supone una novedad en cuanto a su perspectiva, ya que no se centra en obtener un único índice de validez de contenido a partir del juicio de expertos, sino que combina tres índices, ligando la validez de contenido a la validez de constructo para ofrecer una evidencia mucho más exhaustiva.

En este sentido, estos autores proponen calcular la Fiabilidad de Acuerdo Interjueces (IRA, según sus siglas en inglés), el IVC ya definido con anterioridad y el Índice de Validez Factorial (FVI, en inglés).

El índice IRA presenta como finalidad estimar la fiabilidad interjueces derivada del análisis de los ítems en términos de representatividad y claridad del elemento. Para ello, emplean una escala tipo Likert de 4 alternativas que, posteriormente es dicotomizada para seleccionar aquellos ítems considerados adecuados (puntuaciones de 3 y 4 por los expertos). A partir de esta cuantificación, es posible calcular el IRA para cada ítem y para la escala en su conjunto (dividiendo el número de ítems adecuados entre el número total de ítems).

Respecto al IVC, éste índice ya ha sido definido en párrafos precedentes, por lo que, en este caso, implica únicamente una aplicación del índice propuesto por Lawshe (1975).

La novedad de este método surge en el tercer índice a calcular (FVI), el cual aporta información acerca del grado en que los expertos asocian cada ítem con los constructos que se pretenden medir, aportando así una «cuantificación preliminar de la validez factorial» (Rubio et al., 2003, p. 98).

Para calcular el FVI de cada ítem, se divide el número de expertos que asocian correctamente el ítem con su dimensión entre los expertos totales. Este mismo procedimiento, tomando la media del FVI a lo largo de los diferentes ítems puede emplearse para calcular el FVI del instrumento total. A la hora de interpretar el resultado, estos autores proponen alcanzar un valor mínimo de 0.80 para considerar tanto el ítem como la escala adecuados.

Un ejemplo de aplicación de este método, puede consultarse en un interesante trabajo de Yang y Chan (2008) acerca del diseño de páginas web para el aprendizaje de idiomas.

Índice Promediado de la Desviación Media (Claeys, Nève, Tulkens y Spinewine, 2012)

Finalmente, cabe destacar este método, el cual combina el ya conocido IVC con la propuesta de estos autores en torno al Índice Promediado de la Desviación Media (Average Deviation Mean, ADm en inglés).

Bajo este método, en primer lugar, se calcula el IVC de cada ítem y, a continuación, se emplea el ADm para calcular el grado de acuerdo de los expertos independientemente de que estos hayan valorado el ítem positiva o negativamente. Como interpretación, la probabilidad

asociada al ADm deber ser superior al valor crítico de 0.05.

De este modo, ambos índices aportan información complementaria indicando, el primero de ellos, si los expertos aceptan un ítem o no como adecuado y, mediante el ADm, el nivel de acuerdo sobre los citados ítems. En ese mismo trabajo, los autores ejemplifican su método aplicándolo sobre una escala de carácter clínico.

A modo de resumen de los diferentes métodos expuestos, se presenta una síntesis de estos en la Tabla 2.

Tabla 2

Síntesis de los métodos basados en el juicio de expertos para el análisis de la validez de contenido

Año	Autores	Método
1961	Tucker	Basado en Análisis Factorial
1975	Lawshe	CVR
1977	Rovinelli y Hambleton	índice de congruencia ítem-objetivo
1980	Aiken	V
1980, 1984	Hambleton	Índice de congruencia
1986	Lynn	IVC
1992	Sireci y Geisienger	Escalamiento multidimensional y análisis de *clusters*
1997	Levine, et al.	MQ
2001	Fitch, et al.	IPRAS
2002	Hernández-Nieto	CVC
2003	Rubio, et al.	FVI
2012	Claeys, et al.	ADm

Métodos derivados de la aplicación del instrumento de medida

Dejando a un lado el juicio de expertos, existe otra gran perspectiva sustentada sobre metodología estadística. En ella, se alude a procedimientos que analizan los datos obtenidos tras la aplicación de la propia prueba, teniendo en cuenta tanto la puntuación total del test como las respuestas a cada elemento por los participantes evaluados (Sireci, 1998a). Por tanto, la gran diferencia respecto a los métodos previos es que, en este caso, los ítems no son evaluados por un conjunto de expertos, sino que se aplican directamente a un conjunto de participantes para analizar, única y posteriormente, las respuestas dadas por estos. En este sentido, se aludirá a validez de contenido para referirse a la idoneidad de las respuestas dadas por los participantes en relación al constructo que se pretende evaluar, siendo el conjunto de respuestas una muestra del comportamiento de interés (Fitzpatrick, 1983).

Esta perspectiva deriva, precisamente por tener ese carácter más objetivo en donde el participante únicamente debe responder al ítem en base a su conducta, en una importante alternativa a los posibles sesgos que se han apuntado con anterioridad en la valoración de los jueces.

A pesar del importante número de métodos existentes en relación al juicio de expertos, lo cierto es que desde esta perspectiva, los investigadores no se han prodigado tanto a la hora de proponer alternativas que permitan cuantificar la validez de contenido. Aun así, destaca la aplicación específica de pruebas estadísticas que ya se han apuntado previamente como el escalamiento multidimensional y el análisis de clusters, el análisis factorial o la Teoría de la Generalizabilidad.

Si bien es cierto que estas propuestas cuentan con un importante apoyo metodológico, no están, en su mayoría, exentas de limitaciones. Así, aunque el escalamiento multidimensional, el análisis de clusters y el análisis factorial permiten definir claramente los constructos evaluados y su relevancia, su interpretación puede presentar problemas cuando las propiedades de las respuestas obtenidas se solapan con las interpretaciones del contenido (Davison, 1985; Green, 1983).

Una alternativa a esta problemática es la Teoría de la Generalizabilidad (TG). En este procedimiento se diseña, en primer lugar, un estudio de decisión en el que se tienen en cuenta determinadas variables o facetas que constituyen posibles fuentes de error a la hora de analizar la validez (i.e., instrucciones dadas a los participantes, el nivel de habilidad de los participantes, etc.). A continuación se calcula la puntuación media que el conjunto de participantes otorga a todos los ítems. Esto se realiza con el objetivo determinar qué ítems presentan un mayor ajuste al contenido que se quiere evaluar.

Además, teniendo en cuenta los análisis previos, es posible establecer qué facetas son relevantes a la hora de generalizar los resultados del estudio de la validez de contenido. Una aplicación de este método puede consultarse en Crocker, et al. (1988), en donde describen cuatro posibles estudios a la hora de llevar a la práctica la Teoría de la Generalizabilidad.

De los procedimientos anteriores, quizás el empleo del escalamiento multidimensional, precisamente por aportar una visión novedosa, requiera una breve reseña, habiendo definido ya en el apartado precedente el fundamento de los métodos relacionados con el análisis de *clusters* y el análisis factorial.

La lógica que subyace a este método es que aquellos ítems que evalúen constructos similares deben ser percibidos como más próximos entre sí en cuanto a su contenido que aquellos que evalúan cuestiones diferentes. Así, el empleo del escalamiento multidimensional, no sólo permite agrupar estos ítems en torno a un constructo como los métodos precedentes, sino que, además, permite analizar a nivel tanto visual como objetivo el grado de similaridad entre cada uno de estos, respecto a los demás.

Por último, otra de las propuestas que merece especial mención es la planteada por Nunnally y Bernstein (1994). En ella, en primer lugar, establecen la necesidad de calcular la validez convergente de la prueba respecto a un instrumento independiente de la herramienta creada. A continuación, llevan a cabo un estudio de diferencia de medias para analizar el cambio producido en la puntuación obtenida en el constructo de interés tras la aplicación de una intervención determinada. Según esta propuesta, un instrumento contará con una adecuada validez de contenido cuando, además de correlacionar de manera significativa con otra herramienta que evalúe la misma dimensión, se haya producido un cambio por el efecto de una intervención específicamente destinada a modificar dicha variable. Obviamente, a la hora de interpretar los resultados, debe existir una seguridad absoluta en cuanto a la inexistencia de algún tipo de interferencia si el tratamiento es realizado por más de un profesional.

Discusión y conclusiones

Como se ha podido comprobar en la parte inicial del manuscrito, el concepto de validez de contenido ha sido objeto de un largo proceso de modificaciones desde su origen a mediados del S. XX. Sin embargo, estos cambios han estado focalizados en la relevancia que este tipo de validez debe presentar, así como en los diferentes métodos para su estudio, manteniéndose su definición esencialmente estable a lo largo del tiempo.

En este sentido, las posturas en torno al concepto de validez en general han sido diversas y variadas, presentando a lo largo de estas décadas diferentes enfoques, tanto unitarios como fragmentados (Sireci, 2009). Si bien es cierto que han existido (y existen) diferentes perspectivas en este sentido, el acuerdo acerca de la importancia que la validez de contenido presenta a la hora de crear y validar cualquier instrumento de medida es unánime (Abad, Olea, Ponsoda y García, 2011; Kane, 2009).

Dejando a un lado las disquisiciones teóricas, sin duda alguna, los numerosos índices y coeficientes generados a lo largo de los años en torno al estudio de la validez de contenido, revelan la importancia que ésta presenta en el proceso de creación y validación de los instrumentos de medida.

A la hora de decidir qué método emplear en la investigación aplicada, se considera necesario combinar ambas perspectivas, pues como apuntan Haynes, Richard y Kubany (1995), el estudio de la validez de contenido debe ser un proceso multimétodo, tanto a nivel cualitativo como cuantitativo. Un ejemplo de ello es la recomendación de Sireci (1998a), en donde expone cómo el empleo de la Teoría de la Generalizabilidad, unida a la evaluación por parte de los expertos, ofrece un cálculo exhaustivo y preciso de este tipo de validez.

Así, el mero hecho de que contar con un grupo de expertos que informen sobre la falta o exceso de ítems representativos del constructo o que simplemente determinen a qué dimensión corresponde cada elemento, no aporta de por sí información relevante para el proceso de validación (Sireci, 1998a). En este mismo sentido, como indica Fitzpatrick (1983), el uso de métodos únicamente basados en las respuestas dadas por los participantes al test, no garantiza que verdaderamente se esté evaluando la variable de interés a menos que se cuente con evidencias de validez convergente. Por otro lado, si exclusivamente se tienen en cuenta las respuestas al test, esto supone realmente un punto de vista más cercano al estudio de la validez de constructo que de contenido.

Dentro de todos los métodos expuestos en el presente estudio, a nivel aplicado y en relación a los referidos al juicio de expertos, destaca especialmente el uso del IVC planteado por Lawshe (1975). Si bien todos los métodos presentan puntos débiles y críticas, trabajos como el de Polit, Beck y Owen (2007) justifican el empleo de este índice por sus numerosas ventajas, respecto al resto de métodos existentes. Así, en su estudio, estos autores comparan dicho método con un amplio número de índices alternativos y destacan los siguientes beneficios a favor del IVC: facilidad de cálculo, facilidad de interpretación, aporta información tanto a nivel de ítem como de instrumento, así como el hecho de centrar la atención sobre el acuerdo en la relevancia del ítem y el consenso de los expertos más que en la consistencia de las puntuaciones dadas por los jueces.

Aun siendo cierto que el IVC presenta un conjunto de beneficios que señalan su adecuación a la hora de estimar la validez de contenido, se considera el método elaborado por Rubio et al (2003) como uno de los más completos y exhaustivos, si bien, por el contrario, es posible considerar que se extralimita en su cometido, pues vuelve a retomar la ya conocida polémica a la hora de delimitar y relacionar la validez de constructo y contenido (Cronbach y Meehl, 1955; Ebel, 1956).

Respecto al método MQ, se puede señalar la dificultad que entraña aplicar este método a áreas específicas de la Psicología, puesto que el hecho de focalizar la atención en un nivel mínimo de capacidad puede mermar las propiedades psicométricas del instrumento cuando este se aplica a muestras que presentan bajas puntuaciones en la variable evaluada. En este sentido, se entiende complicado poder discriminar entre aquellos sujetos que presentan un nivel bajo en el constructo de interés lo cual, en algunas ocasiones, puede ser especialmente relevante.

De igual manera, aunque el índice CVR presenta una solidez metodológica importante, puede presentar ciertas dificultades de interpretación en algunos casos, pudiendo, por ejemplo, obtener cualquier valor entre ±1 obteniendo, sin embargo, un resultado en CVR = 0 si la mitad de los expertos señalan el ítem como relevante.

Por otro lado, respecto a los métodos derivados de la propia aplicación del instrumento, el empleo del análisis factorial y el análisis de *clusters* presentan una clara orientación hacia la concordancia y la correlación entre los ítems en sí, con lo que se entienden como métodos más cercanos al estudio de la validez de constructo.

Respecto al método expuesto por Nunnally Bernstein (1994), presenta dos claros inconvenientes. En primer lugar, la necesidad de contar

con un instrumento de medida ya validado, lo que implica tanto incrementar la longitud y duración de la aplicación como contar con un instrumento ya existente con adecuadas propiedades psicométricas lo cual, a la hora de trabajar con determinadas variables, puede resultar complicado. Además, este método exige una variación en la variable de medida tras su intervención, por lo que resulta especialmente relevante controlar el efecto de la intervención realizada.

Así pues, se considera que la TG no sólo salva todos estos inconvenientes sino que, como ya se ha señalado, permite, por un lado, determinar qué facetas y en qué medida éstas están afectando a la validez de contenido del test y, además, permite la generalización de los resultados obtenidos siempre que se haya diseño un adecuado estudio de decisión.

Por otro lado, al margen de los métodos para la estimación de la validez y en cuanto al posible sesgo a introducir a la hora de asignar la tarea a los expertos, resulta complicado imaginar que un grupo de expertos evalúe un determinado número de ítems sin conocer realmente qué pretenden medir (Sireci, 2007). Por ello, una de las alternativas es emplear, como señalan Abad et al. (2011), ítems «de relleno», los cuales no miden realmente ninguno de los constructos pero se reduce, de este modo, el citado sesgo. Así, en este punto cobra especial importancia la cantidad y el modo en que la información se ofrece a los expertos para que realicen su tarea, intentando ser lo más asépticos posible en la labor.

Por otro lado, se debe tener presente que el estudio de la validez de contenido no se circunscribe únicamente al análisis de las respuestas o puntuaciones dadas a los ítems, sino que como indican Abad et al. (2011), actualmente existen aspectos que deben ser tenidos en cuenta a la hora de analizar la validez de contenido por presentar un efecto directo en la misma como son la importancia a la hora de aplicar el instrumento y la corrección de la propia prueba.

Como conclusión, se entiende pues que a la hora de estimar adecuadamente la validez de contenido resulta imprescindible la combinación de métodos tanto cualitativos como cuantitativos. Así, una vez construidos los diferentes ítems en torno al constructo a evaluar, sería necesario contar con un grupo de expertos que emitiesen su valoración sobre los mismos. Posteriormente, a la hora de cuantificar la adecuación de dichos ítems, se entiende el método IVC (Lawshe, 1975) como el más adecuado al presentar los mayores beneficios respecto a las diferentes alternativas propuestas a lo largo de los años. Finalmente, una vez definidos qué ítems son relevantes, éstos deberían aplicarse a un conjunto de participantes para, sobre las respuestas dadas por estos, aplicar la TG. Mediante esta metodología sería entonces posible tanto cuantificar el efecto de las posibles fuentes de error, pudiendo así controlarlas en futuras aplicaciones como, principalmente, generalizar los resultados obtenidos.

Agradecimientos: Investigación financiada por el Programa de Formación de Personal Universitario del Ministerio de Educación (AP2010-1999).

Referencias

Abad, F. J., Olea, J., Ponsoda, V. y García, C. (2011). *Medición en ciencias sociales y de la salud [Measurement in Social and Educational Sciences].* Madrid, España: Síntesis.

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement 40*, 955–959.

American Psychological Association. (1952). Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal. *American Psychologist, 7,* 461–465.

American Psychological Association, American Educational Research Association, y National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, y National Council on Measurement in Education. (1999).

Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

Anastasi, A. (1954). *Psychological Testing*. New York: MacMillan.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. En R. L. Thorndike (Ed.), *Educational measurement* (pp. 508–600). Washington, D. C.: American Council on Education.

Balbinotti, M. A. A. (2004). Estou Testando o que Imagino Estar? Reflexoes acerca da Validade dos Testes Psicológicos. En C. E. Vaz y R. L. Graff (Eds.), *Técnicas Projetivas: Produtividade em Pesquisa* (pp. 6-22, 1.ª Ed.). Sao Paulo, Brasil: Casa do Psicólogo.

Balbinotti, M. A. A., Benetti, C. y Terra, P. R. S. (2007). Translation and validation of the Graham-Harvey survey for the Brazilian context. *International Journal of Managerial Finance, 3*, 26-48.

Bazarganipour F., Ziaei S., Montazeri A., Faghihzadeh S. y Frozanfard F. (2012). Psychometric properties of the Iranian version of modified polycystic ovary syndrome health-related quality-of-life questionnaire. *Human Reproduction, 27*, 2729-2736. doi:10.1093/humrep/des199

Buster, M. A., Roth, P. L. y Bobko, P. (2005). A process for content validation of education and experienced - based minimum qualifications: An approach resulting in Federal court approval. *Personnel Psychology, 58*, 771-799.

Castle, N. G. (2008). An instrument to measure job satisfaction of certified nurse assistants. *Applied Nursing Research, 23*, 214-220. doi:10.1016/j.apnr.2008.09.005

Claeys, C., Nève, J., Tulkens, P. M. y Spinewine, A. (2012). Content validity and inter-rater reliability of an instrument to characterize unintentional medication discrepancies. *Drugs Aging, 29*, 577-591.

Crocker, L., Llabre, M. y Miller, M. D. (1988). The generalizability of content validity ratings. *Journal of Educational Measurement 25*, 287-299.

Cronbach, L. J. y Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin 52*, 281–302.

Cureton, E. E. (1951). Validity. En E. F. Lindquist (Ed.), *Educational Measurement* (pp. 621–694). Washington, DC: American Council on Education.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5*, 194-197.

Davison, M. L. (1985). Multidimensional scaling versus components analysis of test intercorrelations. *Psychological Bulletin, 97*, 94–105.

Deville, C. W. (1996). An empirical link of content and construct equivalence. *Applied Psychological Measurement, 20*, 127–139.

Distefano, M. K., Pryer, M. W. y Erffmeyer, R. C. (1983). Application of Content Validity Methods to the Development of a Job-Related Performance Rating Criterion. *Personnel Psychology, 36*(3), 621-631.

Drauden, G. M. y Peterson, N. G. (1974). *A domain sampling approach to job analysis*. Test Validation Center. St. Paul: Minn.

Ebel, R. L. (1956). Obtaining and reporting evidence for content validity. *Educational and Psychological Measurement, 16,* 269–282.

Fitch K., Bernstein S. J., Aguilar, M. D., Burnand, B., LaCalle, J. R., Lazaro, P., … Kahan, J. P. (2001) *The RAND/UCLA Appropriateness Method User's Manual*: RAND corporation.

Fitzpatrick, A. R. (1983). The meaning of content validity. *Applied Psychologyical Measurement, 7*, 3-13.

García-Campayo, J., Zamorano, E., Ruiz, M. A., Pardo, A., Freire, O., Pérez-Páramo. y Rejas, J. (2009). Cultural adaptation into Spanish of the generalized anxiety disorder scale - 7 (GAD-7) scale. *European Psychiatry, 1*(24), 538. doi:10.1016/S0924-9338(09)70771-0

García-Campayo, J., Zamorano, E., Ruiz, M. A., Pardo, A., Pérez-Páramo., López-Gómez, V. y Rejas, J. (2012). Psychometric validation of the spanish version of the GAD-2 scale for screening generalized anxiety disorder. *Health and Quality of Life Outcomes, 19*(10), 114. doi: 10.1186/1477-7525-10-114.

Green, S. B. (1983). Identifiability of spurious factors with linear factor analysis with binary

ítems. *Applied Psychological Measurement, 7*, 3–13.

Guion, R. M. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement 1*, 1–10.

Gulliksen, H. (1950a). Intrinsic validity. *American Psychologist*, *5*, 511–517.

Gulliksen, H. (1950b). *Theory of Mental Tests*. New York: Wiley.

Hambleton, R. K. (1980). Test score validity and standard setting methods. En R. A. Berk (Ed.), *Criterion-Referenced Measurement: The State of the Art*. Johns Hopkins University Press: Baltimore.

Hambleton, R. K. (1984). Validating the test score. En R. A. Berk (Ed.), *A Guide to Criterion-Referenced Test Construction* (pp. 199-230). Baltimore: Johns Hopkins University Press.

Haynes, S. N., Richard, D. C. S. y Kubay, E. S. (1995). Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods. *Psychological Assessment, 7*(3), 238-247.

Hernández-Nieto, R. A. (2002), *Contributions to Statistical Analysis*. Mérida, Venezuela: Universidad de Los Andes.

Kane, M. (2006). Content-related validity evidence in test development. En S. M. Downing y T. M. Haladyna (Ed.), *Handbook of test development* (pp. 131-153). Mahwah, NJ.: Lawrence Erlbaum Associates.

Kane, M. T. (2009). Validating the interpretations and uses of test scores. En R. W. Lissitz (Ed.), *The concept of validity* (pp. 39-64). Charlotte, NC: Information Age Publishing.

Kröger, E., Tourigny, A., Morin, D., Côté, L., Kergoat, M. J., Lebel, P., … Benounissa, Z. (2007). Selecting process quality indicators for the integrated care of vulnerable older adults affected by cognitive impairment or dementia. *BMC Health Services Research, 29*(7), 195. doi:10.1186/1472-6963-7-195

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575.

Levine, E. L., Maye, D. M., Ulm, R. A. y Gordon, T. R. (1997). A methodology for developing and validating minimum qualifications (MQs). *Personnel Psychology, 50*, 1009-1023.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*, 382-385.

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement*, *7*, 191–205.

Muñiz, J. (2000). *Teoría clásica de los tests [Clasical Tests Theory]* (6.ª Ed.). Madrid, España: Pirámide.

Mussio, S. J. y Smith, M. K. (1973). *Content validity: A procedural manual*. Chicago: International Personel Management Association.

Nunnally, J. C. y Bernstein, I. H. (1994). *Psychometric Theory* (3.ª Ed.). New York: McGraw Hill.

Polit, D. F., Beck, C. T. y Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing and Health, 30*(4), 459-467.

Rovinelli, R. J. y Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal of Educational Research, 2*, 49-60.

Rubio, D. M., Berg-Weber, M., Tebb, S. S., Lee, E. S. y Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27*(2), 94-104.

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

Sireci, S. G. (1998a). The construct of content validity. *Social Indicators Research, 45*, 83-117.

Sireci, S. G. (1998b). Gathering and analyzing content validity data. *Educational Assessment, 5*, 299-321.

Sireci, S. G. (2003). Validity content. En R. F. Ballesteros (Ed.), *Encyclopedia of psychological assessment*. Londres, UK: Sage.

Sireci, S. G. (2007). On Validity Theory and Test Validation. *Educational Researcher, 36*(8), 477–481. doi:10.3102/0013189X07311609

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence. En R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and*

applications (pp. 19-37). Charlotte, NC: Information Age Publishing.

Sireci, S. G. y Geisinger, K. F. (1992). Analyzing test content using cluster analysis and multidimensiamal scaling. *Applied Psychological Measurement, 16*, 17-31.

Sireci, S. G. y Geisinger, K. F. (1995). Using subject matter experts to assess content representation: A MDS analysis. *Applied Psychological Measurement, 19*, 241-255.

Merino, C. y Livia, J. (2009). Intervalos de confianza asimétricos para el índice la validez de contenido: Un programa Visual Basic para la V de Aiken [Confidence intervals for the content validity: A Visual Basic com-puter program for the Aiken's V]. *Anales de Psicología, 25*(1), 169-171.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273-286.

Tucker, L. R. (1961). Factor Analysis of Relevance Judgments: An Approach to Content Validity. En A. Anastasi (Ed.), *Testing Problems in Perspective* (pp. 577-586). Washington, DC.: American Council on Education.

Yang, Y-T. C. y Chan, C-Y. (2008). Comprehensive evaluation criteria for English learning websites using expert validity surveys. *Computer and Education, 51*, 403-422. doi:10.1016/j.compedu.2007.05.011

Yeung, E. J. y Shin-Park, K. K. (2006). Verification of the Profile of Mood States-Brief: Cross-Cultural Analysis. *Journal of Clinical Psychology, 62*(9), 1173–1180. doi:10.1002/jclp.20269