Código: PYT-2542-3037-VOL2N2-16-0017

Regresar al Contenido

VALIDEZ Y JUICIO DE EXPERTOS EN LOS INSTRUMENTOS DE INVESTIGACIÓN *

VALIDITY AND JUDGMENT OF EXPERTS IN THE INVESTIGATION INSTRUMENTS

OMAR ESCALONA [1]

RESUMEN

Los que estamos vinculados con el proceso de investigación, hemos tenido que pasar por la validación de los instrumentos para la recolección de información, para lo cual suele apelarse a “la validez y juicio de expertos”. Esta práctica es muy relevante, aunque haya quienes le resten valor. Pero, ¿En qué consiste la validez, sus tipos e importancia?, ¿Cuál es el significado de la validez en una investigación?, ¿Qué es el juicio de expertos?, ¿Cómo es la revisión que hacen los expertos, y qué revisan?, ¿Por qué deben validarse los instrumentos? y, por último, ¿Cuántos expertos validan el instrumento de investigación? ¿Cuál es la diferencia entre jueces y expertos? Estos, entre otros, son algunos de los aspectos desarrollados en el ensayo que aborda la polisemia del término validez y las diferentes implicaciones que tiene el juicio de expertos.

Palabras clave: validez, juicio de expertos, instrumentos de investigación.

ABSTRACT

Those people link with the investigation process, go through the instruments validation to collect information, for what it is used to appealing to "the validity and judgment of experts". This practice is very relevant, although there are those who subtract its value. But, what is the validity, their types and importance? What is the validity meaning in an investigation? What is the judgment of experts? How is the revision experts do? and what do they revise? Why should the instruments be validated? Finally, how many experts validate the investigation instrument? What is the difference between judges and experts? These ones, among others, are some of the aspects developed in the essay that approaches the polysemy of the validity term and the different implications that expert judgment has.

Key words: validity, judgment of experts, investigation instruments.

(*) Enviado: 12-02-2017 Aceptado: 21-03-2017

(1) Ministerio del Poder Popular para la Educación (MPPE)
Email: omarescalona71@gmail.com

INTRODUCCIÓN

El término validez, formulado por Kelly en 1927, es polisémico en su significado. Por lo general, suelen darse tres evidencias con respecto al mismo: *de contenido, de criterio y de constructo* (Babbie, 2014; Hays, 2013; Kellstedt y Whitten, 2013; Hernández, Fernández y Baptista, 2014).

Para el año 1974, la *American Psychological Association (APA)* publicó, en colaboración con *American Educational Research Association (AERA)* y el *National Council on Measurement in Education (NCME)*, que la *validez predictiva y concurrente* se subsumen en *criterial* o *de criterio*. Esta perspectiva de la validez, es denominada por Guion (1980) como la *perspectiva trinitaria* y es la que prevalece en el campo de la psicología (contenido, criterio y constructo).

Sin embargo, en los actuales momentos se habla con frecuencia de otro tipo de validez: la de expertos o *face validity,* que se refiere al grado en que aparentemente un instrumento mide la variable en cuestión, tal como lo han expresado numerosas voces calificadas. Entre estos destacan Gravetter y Forzano (2010), quienes perciben este tipo de validez como otro tipo de evidencia adicional, que por mucho tiempo se consideró parte de la validez de contenido.

En razón de ello, se pretende hacer un recorrido exhaustivo por una decena de autores: Hyrkäs, Appelqvist-Schmidlechner y Oksa (2003), Cabero y Llorente (2013), Corral (2009), Kitamura y Kitamura (2000), Robles y Rojas (2015) y Olaz (2013), entre otros, a fin de develar los planteamientos que aquí se presentan, con la dimensión y amplitud requerida por el tema, clarificando que tal vez sean muchas las inquietudes que se despierten en los lectores acerca de este asunto tan relevante cuando realizan trabajos de investigación.

Por consiguiente, el presente trabajo tiene una gran importancia científica para todo aquel que está pensando en recurrir al Juicio de expertos para validar instrumentos de investigación, y espera generar una reflexión en el lector acerca de lo que significa este aspecto en un trabajo de investigación. La intención no es cerrar la puerta sino dar apertura al espacio de discusión.

DESARROLLO

La validez es un asunto nada sencillo, como lo anunciara Kerlinger (2002), cuando planteó categóricamente la siguiente pregunta al respecto: "¿estamos midiendo lo que creemos que estamos midiendo?" (p.604). Esta interrogante es de suma importancia a considerar cuando se habla de validez de un instrumento de medición. De manera que la validez es una propiedad fundamental, en tanto permite decir de un instrumento que mide lo que pretende medir. En otros términos lo planteó Messick (1995), al describirla como un valor social pero con una perspectiva científica y política. Esto quiere decir que la validez no es cuestión de opciones, de querer o no, sino una exigencia *sine qua non* de los instrumentos de recolección de datos.

Uno de los tipos de validez que se suele aplicar en los trabajos de investigación es la de *criterio de un instrumento.* A propósito, Jackson (2012) refiere que el juicio debe hacerse de forma estándar buscando que exista en el instrumento una relación con los criterios exigidos para su construcción. No obstante, como aclaran Kaplan y Saccuzzo (2013) y Carmines y Zeller (1991), debe tenerse en cuenta la *validez concurrente*, es decir, que el criterio esté presente en el instrumento. A excepción cuando se trata de una *validez predictiva* en la que el criterio está fijado en el futuro.

De acuerdo con los planteamientos de Bostwick y Kyte (2005), para hablar de validez de criterio es necesario que las puntuaciones resultantes en ciertos casos estén correlacionadas y que permitan hacer una predicción de estos mismos casos con respecto a las logradas con otro criterio. Para ello pueden utilizarse el *coeficiente de correlación de Pearson* y el *coeficiente de Spearman*, entre otros. Si los valores difieren o están próximos, indica que los instrumentos utilizados no miden la misma variable, pero sí conceptos relacionados.

Otro tipo de validez muy utilizado en las investigaciones, es la *validez de constructo.* Esta expresión se introdujo por primera vez en 1954 en las *Recomendaciones técnicas para las pruebas psicológicas y las técnicas de diagnóstico.* La *validez de constructo* tiene que ver con la importancia que desempeña la teoría en la elaboración de instrumento a fin de poder plantear hipótesis para someterlas a comprobación o rechazo en el proceso de validación.

En la validez de constructo es necesario contar con varias fuentes de información teórica. Además, se ha fortalecido sobre todo con los trabajos de Cronbach y Meehl, y exige la recogida de diversas evidencias y la integración de información recogida. Así mismo, los trabajos de Campbell y Fiske (1959) aportan conceptos como la *validez discriminante, convergente*, y *la matriz multimétodo-multirasgo* que son un referente en la evaluación de este tipo de validez.

Desde una perspectiva científica, Grinnel, Williams y Unrau (2009), Babbie (2014) y Messick (1975) coinciden en señalar que la validez de constructo es probablemente la más relevante. Para Carmines y Zeller (1979), en la validez de constructo se siguen tres etapas: (a) Establecimiento y especificación entre el concepto o variable medida por el instrumento y los demás conceptos incluidos en la teoría, modelo teórico o hipótesis de acuerdo con la base de la revisión de la literatura. (b) Asociación estadísticamente de los conceptos y análisis cuidadoso de las correlaciones. (c) Interpretación de la evidencia empírica de acuerdo con el nivel en el que se clarifica la validez de constructo de una medición en particular.

El proceso de validación de un constructo está relacionado con la teoría, por tal motivo es vital que exista un marco teórico que avale la variable en correspondencia con otras variables. Desde luego, no amerita un aval teórico muy alto, pero si algo que evidencie la correspondencia entre los conceptos. Por tanto, las preguntas inherentes a la validez de constructo son: ¿El concepto teórico se refleja en el instrumento? ¿Qué expresan las puntuaciones del instrumento? ¿Permite el instrumento medir el constructo y sus dimensiones? ¿Por qué?

Las posiciones de Cronbach (1980), Guion (1980), Linn (1980) y Messick (1980), entre otros, permiten concluir que la *validez de criterio, de contenido, factorial y discriminante* difiere de la validez de constructo y cualquiera de ellas expresa parte de esta.

Por otra parte, también se habla de *validez de expertos* o *face validity*, y está referida al grado en que supuestamente un instrumento mide la variable en referencia, según la voz de diferentes expertos. Asimismo, se vincula con la validez de contenido aunque siempre se le consideró como parte de la misma. Sin embargo, como sostienen Gravetter y Forzano (2010), hoy se le nombra como un tipo adicional de evidencia, y regularmente se establece a través de la evaluación del instrumento ante expertos.

Por su parte, Messick (1995) señala que, la *validez de face* es en sí misma un rasgo deseable de los instrumentos, porque alude a que la prueba "parece valida" para quien la administra, quien la responde y para otros observadores. Recientemente se ha hablado también de la *validez consecuente*, que se refiere a las consecuencias sociales del uso e interpretación de una prueba (Mertens, 2010).

Se precisa además que en un instrumento de medición estén representados todos o la mayoría de los componentes del dominio de contenido de las variables objeto de medición, el cual se puede establecer por la literatura (teoría y trabajos antecedentes). Por esta razón, es adecuado realizar una revisión lo más exhaustiva posible en diferentes estudios, para determinar la dimensión o contenidos en que se ha medido la variable; en caso de ser el dominio de un instrumento muy estrecho con respecto al dominio de las variables, entonces el dominio del instrumento no será representativo. De allí que la *validez de contenido* permite responder a la pregunta: ¿el instrumento realmente mide adecuadamente las dimensiones de la variable?

Fundamentalmente la evidencia de la validez del contenido se obtiene mediante las opiniones de expertos, al revisar que las dimensiones sean medidas por el instrumento y que representen el universo de dimensiones de las variables de estudio, a veces, por muestreo aleatorio simple. En otras palabras, el objeto de estudio debe analizarse sistemáticamente de manera que los ítemes comprendan todos los aspectos necesarios y en la cantidad correcta.

En consecuencia, cuando se hace referencia a la validez, sencillamente se está indicando que el ítem o el instrumento son pertinentes, es decir, miden la variable que deben medir. Porque lo que sí es seguro, es que el lector de la investigación, y en específico del instrumento de medición, cuestione qué tan válida y confiable es la medición.

Muchas veces, en la revisión de trabajos de investigación, se encuentran instrumentos que han sido diseñados en otros países, y esto puede generar ciertas inconsistencias en la medición. Al respecto, Hyrkäs, Appelqvist-Schmidlechner y Oksa (2003) plantean que es muy frecuente, en los países de habla no inglesa, el uso de instrumentos estandarizados de investigación correspondientes a países de habla inglesa, y para ello se recurre a la traducción y adaptación de estos instrumentos. Un ejemplo de estos, es la conocida *The Manchester Clinical Supervision Scale*. La descripción de este método se encuentra en varios documentos. Por ejemplo, *The McGill pain questionnaire*, el McGill pain questionnaire ha sido traducido al noruego, el *Pain Coping Questionnaire* al danés, el *Miller-Rahe Recent Life Change Questionnaire Spanish*, al español, el *Menstrual Distress Questionnaire* al chino y *The Behcet's Disease Current Activity* al turco.

La traducción de escalas algo habitual por parte de sociólogos, psicólogos y educadores. La *International Test Commission (ITC)* encontró que en 1992 algunos tests desarrollados en Estados Unidos fueron traducidos y adaptados a más de 50 lenguas.

Según mencionan Khani, Jaafarpou y Jamshidbeligi (2009), la validación cultural es la fase más importante en el proceso, que puede llevarse a cabo con la prueba piloto y métodos estadísticos. A pesar de ello, es ineludible la evaluación de la validez del instrumento por parte de los expertos, dado que estos son los únicos capaces de realizar la eliminación de aquellos ítems que sean realmente irrelevantes, o simplemente modificar los que sean necesarios cuando se trata de expresiones idiomáticas. Pero, como sustentan Mikulic y Muiños (2004), es obligatorio conocer y tener en cuenta los lineamientos de la ITC sobre todo si se van a realizar estudios transculturales.

Por otra parte, Mikulic (2005) aconseja que los efectos de las diferencias culturales irrelevantes deben minimizarse en los instrumentos de medición. Además, el aspecto crucial al revisar el *background* o antecedentes de los instrumentos a utilizar, son los aspectos lingüísticos del significado de la palabra y el significado de la oración, pero en especial el orden de las palabras y la presencia de giros idiomáticos (Mikulic y Muiños, 2004).

Otro aspecto substancial es la equivalencia métrica, esto es, conocer si los puntajes de las distintas versiones son comparables, lo cual se logra por el *Análisis del Funcionamiento Diferencial de los Ítems* (DIF) y la *detección de los sesgos en los ítems*.

Es importante reseñar que dadas las implicaciones éticas, sociales y jurídicas que envuelve la utilización de tests que pueden infravalorar las capacidades de ciertos grupos en función a su cultura, etnia, sexo o cualquiera otra característica diferenciadora, no es extraño que últimamente una de las áreas más fecundas de la literatura psicométrica sea la dedicada a generar procedimientos para detectar y eliminar aquellos ítems que presentan un funcionamiento diferencial.

Un ítem funcionará diferencialmente (o presenta DIF) cuando dos grupos comparables de sujetos lo ejecuten de manera distinta. Normalmente, el grupo objeto de análisis se denomina *grupo focal* y el grupo que sirve como criterio de comparación se conoce como *grupo de referencia*. Mellenbergh (1982) ha señalado que los ítems de un test pueden presentar

distintos tipos de DIF como: a) *DIF uniforme o consistente* si no existe interacción entre el nivel del atributo medido y la pertenencia a un determinado grupo. b) *DIF no uniforme o inconsistente* cuando ocurre esta interacción, es decir, cuando no hay igual diferencia probabilística en la respuesta al ítem en los dos grupos.

El estadístico de Mantel-Haenszel (MH) es precisamente uno de los que mayor uso tiene a la hora de determinar ítems con DIF, por ser de fácil manejo e interpretación, pues no exige alto dominio de la estadística y tampoco requiere de elevados tamaños muestrales (Mazor, Clauser y Hambleton, 1998). Pero todavía, es un procedimiento poco eficaz para ítems con DIF no uniforme (Hambleton y Rogers, 1989; Hills, 1989).

Entre otros métodos que existen para valorar el DIF están el *análisis de varianza* (Jensen, 1980 y Oestreling, 1983), *el método Delta* (Angoff y Sharon, 1974), *chi-cuadrado* (Camilli, 1979 y Marascuilo, 1981), *modelo loglineales, logit y de clase latente* (Kelderman y Macready, 1990)*, regresión logística* (Spray y Carlson, 1986), *y los métodos basados en teoría de respuesta a los ítems (TRI)* (Kim y Cohen, 1991). Los TRI se determinan con empleo de softwares y por lo general son costosos, además de requerir muestras grandes (Hoover y Cohen, 1984).

Otra de las técnicas que sí controlan el nivel de competencia de los sujetos, son los modelos lineales logarítmicos o logolineales. Autores como Green, Crone y Folk (1989), Mellenbergh (1982), Kelderman y Macready (1990), proponen este tipo de técnicas para detectar el DIF y que permitan formular diferentes hipótesis sobre el tipo DIF o las características de la distribución de las puntuaciones de los sujetos en test.

Del mismo modo, hay que señalar que existen asociaciones que detecta la calidad técnica de los instrumentos. Así por ejemplo, en ciencias específicas, como la psicología, señalan Mikulic y Muiños (2004) que "sería importante que todo psicólogo pudiera disponer de una valoración realizada por personas expertas en el área de evaluación psicológica que informaran la calidad técnica de los instrumentos, construidos y adaptados, en nuestro medio" (p.194). Estos investigadores revelan que en Holanda existe la *Asociación Psicológica Nacional* para revisar los instrumentos que están en el mercado y publicar una guía con toda la información actualizada para los profesionales que utilizan pruebas.

En este mismo orden enuncian Mikulic y Muiños (2004), que en Inglaterra la evaluación de la calidad técnica de los instrumentos la hacen los expertos en de las editoriales. Igualmente sucede en países como España, con la *Comisión de Tests* dentro del *Colegio Oficial de Psicólogos*. También existen organizaciones como la *Federación Europea de Psicólogos* o la *Asociación de Psicólogos Americanos*, que han publicado códigos deontológicos y éticos para la construcción y uso de los instrumentos.

Una pregunta interesante de plantear en este asunto es: ¿cuántos profesionales deben hacer la traducción del test? Respecto de esta interrogante, Hambleton (1994) recomienda un mínimo de cuatro traductores que tengan conocimientos de psicología, pero advierte que la traducción que efectuarla independientemente. Agrega además que es necesario entregarles a los traductores las instrucciones del tipo de traducción a realizar y el tipo de ítems que se espera obtener.

¿Cuántos jueces intervienen en la adaptación del test? En este proceso intervienen un mínimo de cinco jueces bilingües, para evaluar en una escala de cuatro puntos la equivalencia conceptual, en relación al ítem original y el ítem traducido, si es idéntico, bastante similar, bastante diferente y diferente. Posteriormente, dos especialistas evalúan los resultados obtenidos y seleccionan los ítems adaptados. Al mismo tiempo, se analizan los cambios necesarios en los ítems que lo requieran, a objeto de obtener la versión final del instrumento. El otro paso es hacer la prueba piloto por medio de entrevistas semidirigidas a fin de conocer de conocer cuales ítems muestran dificultad en el instrumento. Finalmente se determina el *Alpha de Cronbach* en la

nueva versión y se compara con la versión original para verificar la equivalencia de constructo o hasta qué punto ambos instrumentos evalúan lo mismo en los dos grupos culturales.

La APA afirma que más de 20.000 pruebas específicas nuevas se elaboran cada año así como la revisión de publicaciones anteriores. Hoy día existen diversas organizaciones que han publicado normas de comportamiento ético referidas a la elaboración y uso responsable de pruebas. Las más conocidas según Mikulic (2005) "son las *Standards for Educational and Psychological Testing*, elaboradas por la *Asociación Estadounidense de Investigación Educativa,* la *Asociación Psicológica Estadounidense* y el *Consejo Nacional sobre Medición en Educación"* (p.12).

Las ideas que se han venido expresando delinean otra interrogante: ¿qué es lo que revisan los expertos? A esta pregunta responden Tornimbeni, Pérez, Olaz y Fernández (2004), indicando tres características que deben ser consideradas en cada ítem: (a) Claridad semántica y corrección gramatical, (b) adecuación de su dificultad al nivel educativo y evolutivo de las personas, (c) congruencia con el rasgo o dominio medido.

Por su parte, Messick (1995) alude que entre los aspectos a considerar en la validez están: (a) El contenido. (b) Lo sustantivo. (c) Lo estructural. (d) La generalización. (e) Lo externo. (f) Lo consecuencial.

Respecto del último aspecto, Oesterlind (1990) lo define como el grado de consistencia que debe existir entre el ítem particular y las metas esenciales del instrumento, dado que este será un factor de confiabilidad y validez. A quienes actúan como jueces o expertos se les solicita la evaluación de cada ítem en cuanto a la calidad y consistencia de los mismos, descartándose aquellos ítems que reciben puntuaciones medias bajas y con poco grado de acuerdo entre ellos.

Por su parte, Herrera (1993) recomienda dejar solo los ítems en los cuales el 60 % de los jueces coincidan, así como la necesidad de incluir preguntas que demanden información cualitativa sobre los ítems, lo que puede facilitar un mejoramiento en el posible fracaso de algunos de ellos.

Ahora bien, ¿qué es un experto? Un experto es definido, según Bonano, Hora, Keeney y Von Winter-Feldt (1989), como una persona que posee un conocimiento superior sobre datos, modelos y normas en un área o campo específico.

¿Cuál es la diferencia entre jueces y expertos? Supo (2013) reseña que en la mayoría de los casos se suele considerar estos términos como sinónimos, pero en realidad no lo son. En este sentido, el autor menciona que por ejemplo una mujer de la región alto andina del Perú puede atender un parto en un momento determinado pero no son investigadoras así como tampoco tiene una línea de investigación, pero podría ser investigadora y tener una línea de investigación. En consecuencia, ellas pueden considerarse expertas, pero no podrían ayudar en la evaluación de la idoneidad de los ítems construidos.

Pero si se solicita a un profesional que tiene conocimiento en validación de instrumentos y de técnicas cualitativas o cuantitativas, para evaluar si los ítems han sido redactados correctamente, esa persona es un especialista para evaluar cuestionarios, pero no es un experto en el tema de las costumbres al momento del parto de las mujeres; de esta forma, esta persona puede servir como juez pero no es un experto.

Otra pegunta inherente es: ¿qué requisitos debe tener el experto? Cabero y Llorente (2013) plantean que el término experto es polisémico, y que el modo correcto de aplicación guarda correspondencia con los criterios de selección, del número adecuado de los mismos y de los instrumentos empleados.

Para Hora (2009), es necesario entre esos criterios: (a) Experiencia de investigación probada por las publicaciones y subvenciones. (b) Citas en trabajos. (c) Títulos, premios u otros tipos de reconocimiento. (d) Recomendaciones y nominaciones de órganos y personas respetadas, en cargos desempeñados. (e) Pertenencia para revisar las juntas, comisiones, entre otras. Supo (2013) sostiene la experiencia del investigador es muy importante al

momento de elegir por tanto él es experto y juez dentro de su línea.

Otros autores, como Skjong y Wentworht (2000), proponen los siguientes requisitos: (a) Experiencia en la realización de juicios y toma de decisiones, basada en evidencia o experticia (grados, investigaciones, publicaciones, posición, experiencia y premios entre otras). (b) Reputación en la comunidad. (c) Disponibilidad y motivación para participar. (d) Imparcialidad y cualidades inherentes, como confianza en sí mismo y adaptabilidad. También plantean estos autores que los expertos pueden estar relacionados por educación similar, entrenamiento, experiencia, entre otros; y en este caso, la ganancia de tener muchos expertos disminuye.

En conjunto, Skjong y Wentworht (2000), McGartland, Berg, Tebb, Lee y Rauch (2003) plantean que debe tomarse en cuenta solo el número de publicaciones o la práctica comprobada. Asimismo, Cook (1991) sostiene que los expertos tienen que estar dispuestos a participar y ser responsables de los juicios emitidos.

Desde otra perspectiva, sostiene Hora (2009), es necesario considerar que los expertos deben contar con algunos requisitos adicionales como por ejemplo, la decisión debe estar libre de sesgos provocados o debida a un interés económico, político u otro. Esto significa estar dispuestos a que sus nombres estén asociados con sus respuestas específicas. A veces, la proximidad física o disponibilidad es una consideración importante.

Pero, ¿cuántos expertos deben seleccionarse? Las experiencias han demostrado que las diferencias entre los expertos pueden ser muy importantes en la determinación de la incertidumbre total expresada en una pregunta. Además, el número de jueces que se debe emplear en un juicio depende del nivel de experticia y de la diversidad del conocimiento. Es conveniente aclarar que es falso el hecho que debe ser un número impar, dado que el juicio es una apreciación cualitativa y no persigue como fin hacer determinaciones estadísticas.

En este orden ideas, Clemen y Winkler (1985), llegó a la conclusión que son adecuados de tres a cinco expertos. Por su parte, Hora (2004) recomendó de tres a seis o siete expertos como suficientes, esperándose pocos beneficios adicionales más allá de ese punto. Del mismo modo, McGartland *et al.* (2003) sugieren un rango de dos hasta veinte expertos. Por su parte, Hyrkäs *et al.* (2003) expreso que deben ser diez y sugieren que para incluir un ítem en un instrumento la coincidencia de validez entre los expertos debe de 80%.

En relación al juicio de expertos, ¿Cuáles son las ventajas y campos en que se utiliza para emitir a validez? Escobar-Pérez y Cuervo-Martínez (2008) definen el juicio de expertos como “una opinión informada de personas con trayectoria en el tema, que son reconocidas por otros como expertos cualificados en éste, y que pueden dar información, evidencia, juicios y valoraciones” (p.29). También se le admite como una práctica que “consiste, básicamente, en solicitar a una serie de personas la demanda de un juicio hacia un objeto, un instrumento, un material de enseñanza, o su opinión respecto a un aspecto concreto” (Cabero y Llorente, 2013 p. 14). Mediante el juicio de expertos se pretende tener estimaciones razonablemente buenas, las “mejores conjeturas” (Corral, 2009 p.231). No obstante, hay que aclarar que tales estimaciones pueden cambiar en el transcurso del tiempo, si se dispone de nuevos fundamentos teóricos respecto al tema.

Así mismo, hay que mencionar que existen ventajas al usar la evaluación por juicio de expertos, como estrategia; entre ellas se tienen la calidad teórica de la respuesta que se obtenida de la persona, la profundidad de la valoración ofrecida, la facilidad de ponerla en acción, pocos requisitos técnicos y humanos para su ejecución, la disponibilidad de uso de diferentes estrategias para recabar sobre contenidos y temáticas difíciles, complejas y novedosas o poco estudiadas, y la posibilidad de obtener información detallada sobre el tema estudiado con participación de diferentes tipos de expertos (Cabero, 2001; Barroso y Cabero, 2010).

Se utiliza en campos como: investigación científica, neurología, psicología, psiquiatría y educación. Un ejemplo es el que mencionan Kitamura y Kitamura (2000), investigadores del *Instituto Nacional de Salud Mental y del Centro Nacional de Neurología y Psiquiatría*, de Ichikawa, Japón, quienes reseñan un estudio con 176 miembros de la *Sociedad Japonesa de Psiquiatría y Neurología*, y en el que se emitió un juicio clínico a través de un cuestionario de competencia en pacientes psiquiátricos, para dar consentimiento informado. Igualmente, Olea, Abad y Ponsoda (2002), de la Universidad Autónoma de Madrid, comentan la manera como se redactó un banco de 635 ítems de gramática inglesa e informan cómo se efectuó diseño de anclaje para la calibración del banco, incidiendo en la predicción de dificultad de los ítems a partir de las valoraciones de expertos.

Igualmente, Robles y Rojas (2015), de la Sapienza Università di Roma, han reseñado las implicaciones surgidas al poner en práctica una validación basada en el juicio de expertos, proporcionando ejemplos de su implementación en investigaciones cualitativas en el campo de Lingüística Aplicada a la enseñanza de lenguas. Se trata de dos investigaciones desarrolladas en el contexto académico con utilización de instrumentos de recogida de información distintos y pertinentemente validados en cada caso.

Por último, hay que mencionar que existen múltiples formas para obtener el juicio de expertos, desde sencillas, hasta las más complejas en su nivel de estructuración. Los expertos como Corral (2009), Robles y Rojas (2015) mencionan la existencia de métodos individuales y grupales o colectivos entre los que destacan: *el método individual o de agregados individuales, el método Delphi, la técnica nominal y el método de consenso grupal*).

CONCLUSIONES

A partir de la revisión bibliográfica hecha queda clara la polisemia del término validez (Babbie, 2014; Kellstedt y Whitten, 2013 y Guion, 1980). Se ha dado a conocer que los validadores, para emitir el juicio de expertos, deben cumplir diversos requisitos como profesionales, junto al número indicado y los diferentes criterios a considerar al momento de realizar la validación, que varían de una universidad o centro de investigación a otro, así como del tipo de investigación que se realiza. Vale decir que no existe un patrón o estándar único de impacto para hacer esa revisión (Hora, 2009; Skjong y Wentwortht, 2000; McGartland, Berg, Tebb, Lee y Rauch, 2003). Pero sí resulta aconsejable, por experiencia profesional vivida como investigador, que antes de aplicar un instrumento de medición se preste especial cuidado a cada observación que emane del experto, dado que el instrumento debe medir la variable con el número de ítems e indicadores que sean necesarios y no otra cosa. Además, posterior a la validación, cuando se realiza la prueba piloto, se devela a través de la determinación del valor de la confiabilidad que ese instrumento de investigación *per se* cumple dos condiciones básicas: la validez y la confiabilidad.

La validez y el juicio de expertos determinan que el instrumento de medición esté diseñado con rigurosidad científica, a objeto de obtener resultados validos que permitan tomar decisiones o hacer el análisis de resultados porque los instrumentos de medida son en sí técnicas a través de las cuales, por codificaciones numéricas, se establece la manera cómo se expresa un constructo, que además sigue un paradigma, una perspectiva teórica, un diseño metodológico y un análisis de datos, por ello se precisa el juicio de expertos para su validez.

Finalmente, se espera haber despertado el interés y el sentido de responsabilidad que se tiene que asumir cuando se realiza una investigación científica, desde esa cuidadosa revisión de literatura, para poder formular las dimensiones de contenido más pertinentes que permitan medir las variables; aspectos tratados en este ensayo. Aunque existen muchas necesidades apremiantes cuando se realiza un trabajo de investigación, el juicio de expertos debe ser punto de agenda del trabajo ineludible en el proceso de investigación a desarrollar.

REFERENCIAS

Angoff, W. H. y Sharon, A. L. (1974). The evaluation of differences in test perfomnance of two or more groups. Journal of Educationa Measurement, 10, 95-105.

Babbie, E. R. (2014). The Practice of Social Research (14th Edition ed.). Springfield, VA, USA: Rawat Publications.

Barroso, J. y Cabero, J. (2010). La investigación educativa en TIC. Madrid, España: Síntesis.

Bernard, H. R. (1988). Research methods in cultural anthropology. Sage, Newbury Park, USA.

Bonano, E. j., Hora, S. C., Keeney, R. L. y Von Witer-Feldt, D. (1989). Elicitacion and use of expert judgmentin performance assessment for high-level radioactive wastw repositories. Washington, DC: NUREG/CR-5411, U.S. Nuclear Regulatory Commission.

Bostwick, G. J. y Kyte, N. S. (2005). Measurement. In E. M. Grinnell, Social work: Research evaluation. Quantitative and qualitative approaches (7a ed., pp. 97-111). New York, NY, EE.UU: Oxford University Press.

Cabero, A. J. y Llorente, C. M. C. (2003, Julio - Diciembre). La aplicación del juicio de experto como técnica de evaluación de las tecnologías de la información (TIC). Revista de Tecnología de Información y Comunicación en Educación, 7(22), 11-22. Retrieved Julio 11, 2016, from http://servicio.bc.uc.edu.ve/educ/eduweb/v7n2/art01.pdf

Camilli, G. (1979). A critique of the chi-square method of assessing item bias. Laboratory of Educational Research.

Campbell, D. T. y Fiske, A. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Carmines, E. G y Zeller, R. A. (1979). Realiability and validity assessment. Beverly Hills, California, United States of America: SAGE Publications, Inc.

Clemen, R. T. y Winkler, R. L. (1985, Mar-Apr). Limits for the precision and value of information from dependent sources. Operations Research, 33(2), 427-442. Retrieved Julio 07, 2016, from https://faculty.fuqua.duke.edu/~clemen/bio/Published%20Paperspdf

Cooke, R. M. (1991). Experts in unicertainty. Oxford: Oxford University Press.

Corral, Y. (2009, Enero - Junio). Validez y confiabilidad de los instrumentos de investigación para la recolección de datos. Revista de Ciencias de la Educación, 19(33), 228 - 247. Retrieved Julio 11, 2016, from http://servicio.bc.uc.edu.ve/educ/revista/n33/art12.pdf

Delbecq, A., Van de Ven, A. y Gustafson, D. (1975). Group Techniques for Program Planning. Glenview, IL: Scott, Foresman y Co.

Escobar-Pérez, J. y Cuervo-Martínez, Á. (2008). Validez de contenido y juicio de expertos: Una aproximación a su utilización. Avances en Medición, 6, 27-36. Retrieved Julio 08, 2016, from http://www.humanas.unal. edu.co/psicometria /files/7113/8574/5708/Articulo3_Juicio_de_expertos_27-36.pdf

Gravetter, F. J. Forzano, Lori-Ann. B. (2010). Research Methods for th Behavioral Sciences (4th ed.). USA: CENGAGE Learnin Customer.

Grinnell, R. W. (2009). Research methods for BSW students (Senenth edition ed.). Mishawaka: Pair Bod Publishing.

Guion, R. M. (1980). On trinitarian doctrines of validity. Professional Psychology, 11, 385-389.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: a progress report. European Journal of Psychological Assesmente, 10, 229-244. Retrieved Julio 5, 2016, from http://files.eric.ed.gov/fulltext/ED399291.pdf

Hambleton, R. K. y Roger, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area an Mantel-Haenszel methods. Applied Measurement in Edacation, 2(4), 313-334.

Hernández, S. R., Fernández, C. C., y Baptista, L. P. (2014). Metodología de la Investigación (6a ed.). México, México: McGraw Hill.

Herrera, R. A. (1993). La medición en psicología. Bogotá: Universidad de Bogotá.

Hora, S. C. (2004). Probability judgments for continuous quantities: linear combinations and calibration. Management Science, 50, 597-604. Retrieved Julio 10, 2016, from http://dx.doi.org/7/mnsc.1040.0205

Hora, S. C. (2009). Expert judgment in risk analysis. Retrieved Julio 07, 2016, from Non-published Research Reports. Paper 120: http://research.create.usc.edu/nonpublished_reports/120?utm_source=research.create.usc.edu%2Fnonpublished_reports%2F120yutm_medium=PDFyutm_campaign=PDFCoverPages

Jackson, S. L. (2012). Research Methods and Statistic: A critical thinking approach (5a ed.). Boston, USA: CENGAGE Learning.

Kaplan, R. M. y Saccuzzo, D. P. (2013). Psychological testing: Principles, applications, and issues (8th ed.). USA: CENGAGE Learning.

Kelderman, H. y. (1990). The use of loglinear models for assessing differential item functionig across manifest and latent examine gruops. Journal of Educational Measurement, 22, 307-327.

Kellstedt, P. M. y Whiten, G. D. (2013). The fundamentals of political science research (2 Edition ed.). Texas.

Kerlinger, F. N. y Lee, H. B. (2002). Investigación del comportamiento: Métodos de investigación en ciencias sociales. 4ª edición. México, DF, México: McGraw-Hill Interamericana Editores.

Khani, A., Jaafarpou, M. y Jamshidbeligi, Y. (2009). Translating And Validating The Iranian Version Of The Manchester Clinical Supervision Scale (MCSS). Journal of Clinical and Diagnostic Research, 3, 1402-1407. Retrieved Julio 08, 2016, from http://www.jcdr.net/back_issues. asp?issn=0973-709xyyear=2009ymonth=Aprilyvolume=3yissue=2y page =1402-1407yid=399

Kim, S. y Cohen, A. S. (1991). A comparison of two areas measures for detecting differencial item functioning. Applied Psychological Meassurement, 22, 269-278.

Kitamura, T. y Kitamura, F. (2000, April). Reliability of clinical judgment of patients' competency to give informed consent: A case vignette study. Psychiatry and Clinical Neurosciences, 54(2), 245-247. doi:DOI: 10.1046/j.1440-1819.2000.00665.x

Linn, R. L. (1980). Issues of validity for criterion-referenced measures. Applied Psychological Measurement, 4, 547-561.

Marascuilo, L. A. y Slaugther. (1981). Statistical procedures for identifying possibl sorces of items bias based on chi-square stastics. Journal of Educational Measurement, 18, 229-248.

Mazor, K. M., Clauser, B. E. y Hambleton, R. K. (1992). The efect of sample size on te functioning of the Mantel-Haenszel statistic. Educational and Psychological Measurement, 52, 443-451.

McGartland, D; Berg, M; Teb , S; Lee, S. S; Rauch, S;. (2003). Objectifying content validity: Conducting a content validity study in social work research. Social Work Research, 27(2), 94-104. Retrieved Julio 09, 2016, from http://swr.oxfordjournals.abstract

Mellenbergh, G. J. (1982). Contingency table models for assesing item bias. Journal of Educational Statistic, 7, 105-118.

Mertens, D. M. (2010). Research and evaluation in education and psychology: integrating diversity with quantitative, qualitative, and mixed methods (3rd ed.). Los Angeles: SAGE Publications, INC.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012-1027.

Messick, S. (1995). Validity of psycological assessment. American Psychologist, 741-749. Retrieved Julio 09, 2016, from http://www.radford Sprin2007/ messick validities.pdf

Mikulic, I. M. (2005). Construcción y adaptación de pruebas pscicológicas. Buenos Aires: Universidad de Buenos Aires.

Mikulic, I. M. y Muiños, R. (2004). La construcción y uso de instrumentos de evaluación en la investigación e intervención psicológica: El inventario de calidad de vida percibida (ICV). XII Anuario de Investigaciones, 193-202. Retrieved Julio 07, 2016, from http://www.redalyc.org/articulo. oa?id=369139941019

Oesterlind, S. (1990). Establishing criteria for meritorius test ítems. Educational Resarch Quality, 14(3), 26-30.

Olaz, C. A. J. (2013). La técnica de grupo nominal como herramienta de investigación. Revista de la Asociación de Sociología de la Educación (1), 114 - 121.

Olea, J., Abad, F. J. y Ponsoda, V. (2002). Elaboración de un banco de ítems, predicción de la dificultad y diseño de anclaje. Metodología de las ciencias del comportamiento, Volumen Especial, 427-430. Retrieved Julio 10, 2016, from http://www.uam.es/personal_pdi/psicologia/fjabad/cv/articulos/ARTICULOSrmcc/mso45657.pdf

Robles, G. P. y Rojas, M. D. C. (2015). La validación por juicio de expertos: dos investigaciones cualitativas en Lingüística aplicada. Revista Nebrija de Lingüística Aplicada (2015), 18.

Spray, J. y Carlson, J. (1986). Comparison of loglinear and logistic regression model for detecting changes in proportions. Paper present at annual meetin of th American Educational Reserach Association. San Francisco.

Supo, J. (2013). Cómo validar un instrumento. La guía para validar un instrumento en 10 pasos. Lima: www.bioestadístico.com.

Tornimbeni, S., Pérez, E., Olaz, F. y Fernández, A. (2004). Introducción a los tests psicológicos (3a edición Revisada y aumentada ed.). Buenos Aires, Córdoba, Argentina: Editorial Brujas.

(*) Enviado: 29-01-2017 Aceptado: 12-03-2017

(1) Ministerio del Poder Popular para la Agricultura Urbana - Ciara (MPPAU-CIARA)
Email: karicaro@gmail.com