VALIDEZ Y CONFIABILIDAD: CRITERIOS FUNDAMENTALES DE CALIDAD MÉTRICA

*Elsy Urdaneta Durán**

RESUMEN

En esta investigación se hace una revisión acerca de los dos principales criterios de calidad métrica que debe satisfacer toda investigación científica: la validez y la confiabilidad. Se revisan las publicaciones más importantes que guían la construcción de tests para medir variables del campo psicosocial y se describe la evolución de los conceptos de validez y confiabilidad tanto histórica como metodológicamente, explicando los procedimientos que permiten recoger evidencias para argumentar la validez de las interpretaciones y estimar la confiabilidad de las puntuaciones obtenidas al aplicar un instrumento para medir variables del ámbito psicológico, social o educativo.

Palabras clave: *validez y confiabilidad, calidad métrica.*

* Doctora en Metodología de las Ciencias del Comportamiento (Universidad Autónoma de Madrid), Diplomado de Estudios Avanzados en Metodología de las Ciencias del Comportamiento (UAM). Profesora Asociada en el Área de Estadística en el Núcleo "Rafael Rangel" de la Universidad de Los Andes Trujillo–Venezuela, E-mail: elsyurdaneta@ula.ve

Recibido:12/09/2013 **Aprobado:** 03/12/2013

VALIDITY AND RELIABILITY: KEY CRITERIA OF METRIC QUALITY

ABSTRACT

This research is a review about the two main criteria of quality metrics that should satisfy any scientific research: validity and reliability. We review the most important publications that guide the construction of tests for measuring variables of the psychosocial field and describes the evolution of the concepts of validity and reliability both historically and methodologically, explaining the procedures that allow gather evidence to argue the validity of interpretations and estimate the reliability of the scores obtained by applying an instrument to measure variables of psychological, social or educational field.

Key Word: *validity and reliability, metric quality.*

INTRODUCCIÓN

Los avances científicos se orientan hacia la mejora de la vida del hombre y de su entorno; así los científicos se plantean como finalidad el conocimiento de la realidad a objeto de poder sistematizarla y de esa manera predecir, tomar decisiones y contribuir al bienestar. Para ello, lo primero es describir los diversos objetos de estudio y para eso se necesita a su vez disponer de procedimientos de medición.

Considerando que los métodos de la ciencia involucran equipamiento sofisticado, controles estrictos y procedimientos altamente estandarizados que tratan de garantizar la objetividad y comparabilidad de las medidas obtenidas, se ha aceptado que, aunque los tests o cuestionarios no son lo mismo que los instrumentos de medición utilizados en las ciencias naturales, se han desarrollado importantes teorías que justifican y avalan su uso, sobre la base de la interpretación que puede hacerse de las puntuaciones obtenidas con ellos. Dado que los tests han sido y siguen siendo herramientas de gran utilidad ampliamente empleados en la medición de las variables del ámbito social y psicológico,

ha sido necesario investigar en profundidad todo lo relacionado con la elaboración de los ítems y el funcionamiento de las puntuaciones obtenidas con ellos.

El hecho de que se pueda extraer información que permita hacer predicciones y tomar decisiones lo más sólidamente fundadas acerca de los fenómenos estudiados, justifica la necesidad de la investigación psicométrica, que se consolida como un área de gran interés en el contexto social y una tarea fundamental dentro de la investigación científica contemporánea. Los avances que se hagan en el perfeccionamiento de la medida de constructos psicológicos, educatios y sociales permitirán el diseño y construcción de instrumentos que aporten puntuaciones con una adecuada calidad métrica que hagan posible la interpretación y utilización de las mismas para el fin previsto.

Pero independientemente de los numerosos estudios acerca de la medida de constructos psicológicos y sociales, hay un par de preguntas persistentes que siempre deben plantearse en relación con los tests: ¿qué miden? y ¿qué tan bien lo miden?

La primera pregunta ha sido objeto de numerosas investigaciones en el ámbito propiamente psicométrico y se identifica con el problema de la validez de las inferencias realizadas a partir de las puntuaciones en el test. La segunda pregunta se asocia con el estudio de la confiabilidad de estas puntuaciones, es decir, de la precisión de las medidas obtenidas con los tests.

La validez y la confiabilidad representan los dos criterios fundamentales de calidad de la medida de una variable a investigar científicamente. Esta es la razón por la cual estos criterios deben ser satisfechos en toda investigación y por ende la actualización sobre estos tópicos es absolutamente necesaria. Para esto, tal como señala Frisbie (2005), es necesario acudir a las fuentes de consenso en las cuales confía la psicometría. Las principales referencias en relación a conceptos, significados de términos especializados y expectativas para una práctica aceptable son la versión actual de los *Standards for educational and psychological testing*(American Educational Research Association, American Psychological Association y National Council on Measure-

ment in Education, 1999) y el manual *Educational Measurement*, tanto en la versión editada por Brennan (2006), como en la de Linn (1989), donde se encuentran las definiciones y normativas autorizadas en lo que a medición respecta.

Los estudios de validez, afirman Abad, Olea, Ponsoda y García (2011), son los que aportan el significado a las puntuaciones que se obtienen al aplicar un test. Pero, tal como señala Frisbie (2005), un concepto fundacional de los aspectos más importantes de la medidacontinúa siendo uno de los peor comprendidos y utilizados. Pese a que ya en los estándares sobre tests del año 1999 (American EducationalResearch Association, American Psychological Association y National Council on Measurement in Education) se señala que la validez esta relacionada con la interpretación que se de a las puntuaciones, aun hoy en día se pueden leer artículos de revistas científicas y libros sobre medición y metodología que se refieren a la validez como una característica del instrumento de medición.

En tal sentido, Frisbie (2005) insiste en que hablar de la "validez del instrumento" puede generar distorsiones metodológicas y plantearía preguntas como las siguientes: ¿puede medir el instrumento un constructo diferente al que plantea el constructor del mismo?, ¿puede un instrumento válido producir puntuaciones que no representan el significado que planteó el constructor del instrumento? La respuesta a ambas preguntas es afirmativa, se puede obtener información equivocada con un buen instrumento. Por ejemplo, si los encuestados responden sin tomarse en serio el test, las puntuaciones no son válidas. Si el encuestado encuentra mejor copiarse las respuestas del vecino, las puntuaciones no son válidas. Si las puntuaciones están analizadas con una técnica estadística para la cual no se cumplen los supuestos de aplicación, las inferencias no son válidas. Si el tiempo no es suficiente para contestar el cuestionario es probable que la falta de respuesta se deba a esto y no a otro factor, por lo cual el resultado de las mediciones no es válido. La persona que digitalizó los datos no atendió a la codificación estipulada, por lo tanto, los resultados no son válidos. Y todas estas situaciones son independientes del contenido del test. En consecuencia, es arriesgado afirmar que un instrumento bien elaborado, en acuerdo con su matriz de contenidos o mapa de variables, con la redacción adecuada, el formato

de respuesta conveniente y todas las consideraciones técnicas que se precisan, sea un instrumento que siempre aporte inferencias válidas. En razón de lo anterior, lo que interesa que sea válido no es el instrumento, sino las interpretaciones que se hagan de las puntuaciones obtenidas con ese instrumento.

De la misma manera, la confiabilidad también es de las puntuaciones obtenidas y no del instrumento de medición. Esto viene como consecuencia de lo anteriormente planteado, si las puntuaciones obtenidas permiten una interpretación válida en cuanto que su significado se corresponde con el de la variable a medir, la precisión de la medida estará asociada a las puntuaciones y no al instrumento con que fueron hechas. Adicionalmente, al ser una característica dependiente del conjunto de ítems y del grupo de sujetos examinados, las estimaciones de la confiabilidad siempre van a variar al cambiar el contexto de aplicación y la muestra utilizada, de modo que el mismo instrumento puede arrojar estimaciones de confiabilidad diferentes si es aplicado en situaciones distintas.

Aclarados estos importantísimos extremos conceptuales se hará un recorrido histórico de la definición de validez hasta detenernos en la concepción actual, explicando las fuentes de evidencias necesarias para argumentar la validez de las inferencias. En relación a la confiabilidad, se explicará cómo es el tratamiento de la misma en función de la teoría psicométrica utilizada y se puntualizará acerca de los principales procedimientos utilizados para la estimación de la misma.

LA VALIDEZ DE LAS INFERENCIAS

Aun cuando el concepto de validez ha evolucionado enormemente, su consideración como la propiedad fundamental de la medida se ha mantenido a lo largo de la historia de la psicometría (Abad et al. 2011).

La definición que hace Messick (1989) del término validez en la tercera edición del manual *Educational Measurement* señala que es un juicio evaluativo integrado del grado en el cual la evidencia empírica y los razonamientos teóricos dan soporte acerca de la adecuación y propiedad de las acciones e inferencias basadas en las puntuaciones

de los tests u otros modos de evaluación. En los *Standards for educational and psychological testing*(AERA, APA, NCME; 1999) se define validez como el grado en que la teoría y los datos disponibles apoyan la interpretación de las puntuaciones obtenidas de un cuestionario para el uso que fue proyectado.

Para llegar a estas definiciones la noción de validez ha ido evolucionando a lo largo de su historia tanto desde el punto de vista de su significado (como se ha podido podido ver) como el de su tratamiento.

Se parte de una concepción empirista que considera la validez como una propiedad del test que puede ser analizada en base a su correlación con un criterio externo. Sin embargo, esta concepción de validez no era adecuada para cubrir todo lo que demandaba el ámbito escolar, en el cual desde un enfoque operacional, las evidencias referidas al contenido son muy importantes y pueden entonces ser consideradas como fuente de argumentos de validez, sostenidos sobre la base de que los ítems del test han de ser representativos de la variable que pretende medir. Con el auge de las teorías factorialistas de la inteligencia, la validez comienza a valorarse desde una perspectiva menos ateórica donde se le considera como el grado en que las puntuaciones del test tienen un significado psicológico en correspondencia con el constructo que se desea medir; aparece entonces en escena la validez de constructo, que fuera incluida como tal en la primera versión de los *Standards for educational and psychological testing*(APA, 1954).

Es así que el concepto de validez ha ido moviéndose hacia una noción más amplia en la cual lo que se valida son las inferencias que se hagan a través de medidas obtenidas con el instrumento y el proceso de recogida de evidencias de validez no se centra en el cuestionario, sino en el uso y la interpretación de las puntuaciones. El análisis de la validez se asume desde una perspectiva integradora que considera la validez como un proceso mediante el cual el investigador debe reunir un conjunto de evidencias argumentativas, para poder sostener la afirmación de que las inferencias que se hagan basadas en las puntuaciones son válidas. Tanto es así que en la cuarta edición del manual *Educational Measurement,* Kane (2006) etiqueta su capítulo sobre el tema como *Validation* y no como *Validity*. Concebida de esta manera, la validez es un proceso de

recolección de evidencias y resulta claro que para obtener estas evidencias se puede usar una enorme variedad de métodos y estrategias (Abad, Olea, Ponsoda y García; 2011). Desde esta perspectiva, se entiende que no debemos referirnos a "tipos de validez" sino a "tipos de evidencia".

Los dos últimas ediciones de los *Standards foreducational and psychological testing*(1985, 1999) recogen esta perspectiva integradora del concepto de validez y en la última edición se establecen cinco fuentes de evidencias para argumentar la validez de las inferencias, estas son las basadas en el contenido del test, en la estructura interna del test, en las relaciones con otras variables, en los procesos de respuesta y en las consecuencias del test. Señalan también que los argumentos de validez que se logren reunir deben articularse de modo tal que puedan respaldar las interpretaciones que se hagan de las puntuaciones dentro del contexto en el cual se hace uso del test, dando recomendaciones operativas específicas acerca de datos empíricos o argumentaciones teóricas que deben aportarse en determinadas situaciones.

Evidencias basadas en el contenido del test

Hay dos aspectos básicos a considerar en la recogida de evidencias de validez del contenido: la definición de la variable y la representación de la misma.

Definición de la variable

Se relaciona con la operacionalización de la misma. Comúnmente esta actividad se concreta mediante la tabla de especificaciones o el mapa de variable. En la tabla de especificaciones se construye una matriz detallada que contiene todos los aspectos relacionados con ese constructo o variable y se usa especialmente en pruebas cognitivas. En el mapa de variables se operativizan los objetivos y se categoriza la variable en dimensiones y subdimensiones (si las hubiere), usándose esta estrategia principalmente en cuestionarios de carácter no cognitivo.

Representación de la variable

Abarca dos aspectos: representatividad y relevancia. La representatividad o cobertura indica la adecuación con que el contenido del instrumento asume todas las facetas o dimensiones de la variable, examinando si todas las dimensiones están siendo consideradas o si hay algunas que están infrarrepresentadas. La relevancia indica el grado en que cada ítem mide la variable definida, pudiéndose detectar problemas relativos a la presencia de contenidos irrelevantes.

La mayoría de estudios de validación de contenidos requieren el juicio de expertos o jueces que evalúen los ítems del instrumento y emitan valoraciones sobre el grado de emparejamiento entre los ítems y los objetivos enunciados en la tabla de especificaciones.

Evidencias basadas en la estructura interna del test

Permiten argumentar que el instrumento es un constructo coherente y no es simplemente un conjunto de ítems no relacionados. Para analizar la estructura interna del test se realizan estudios sobre la dimensionalidad y el funcionamiento diferencial de los ítems.

Dimensionalidad

En estos estudios se determina si la estructura del instrumento coincide con la estructura teórica postulada al construir el cuestionario. Para esto se utiliza principalmente el análisis factorial que examina si los ítems se agrupan en factores que representan las dimensiones planteadas en la tabla de especificaciones o el mapa de variables.

Funcionamiento diferencial del ítem

Se presenta cuando personas con el mismo nivel en la característica medida, pero pertenecientes a grupos distintos, tienen distinta probabilidad de estar de acuerdo con el ítem. Las técnicas más utilizadas para detectar funcionamiento diferencial son la de Mantel-Haenszel y la regresión logística. La existencia de funcionamiento diferencial es evidencia de problemas en el instrumento que acarrearan inferencias con escasa validez.

Evidencias basadas en la relación con otras variables

El objetivo de este procedimiento es establecer si las relaciones observadas entre las puntuaciones obtenidas con el cuestionario y otras variables externas que se sabe relevantes al constructo son consistentes con la interpretación propuesta. Las variables externas pueden ser: a) otras medidas del mismo constructo obtenida con otro instrumento (evidencia convergente), b) medidas de un constructo diferente que se inserta en el modelo teórico donde se encuadra la variable de interés (evidencia discriminante) y c) algún tipo de variable que pretendamos predecir a partir de las puntuaciones del test o que se sepa por la teoría existente que está fuertemente correlacionada con la variable que mide el instrumento (validez referida a un criterio).

Evidencia convergente y discriminante

Para la obtención de evidencia convergente y discriminante se busca examinar las relaciones previsibles entre las puntuaciones obtenidas con el cuestionario y otras variables, ya sean similares (evidencia convergente) o diferentes (evidencia discriminante). El procedimiento más adecuado para hacer esto es la estimación y análisis de la matriz multirasgo multimétodo para la cual es necesario seleccionar las variables (rasgos) y los métodos de modo que cada uno de los métodos sea adecuado para medir todas las variables de interés, los diferentes métodos sean independientes entre si y las variables incluidas varíen en el grado de asociación entre ellas, con unas altamente relacionadas y otras cuya asociación sea muy baja. La interpretación exhaustiva puede encontrarse en Abad et. al. (2011), no obstante, de manera directa se puede afirmar que existen evidencias convergentes cuando variables similares dan correlaciones altas con el mismo método y evidencias discriminantes cuando variables diferentes medidas con distinto métodos tienen correlaciones que tienden a cero.

Validez referida a un criterio

El procedimiento a seguir para obtener evidencias referidas a un criterio, representado por una variable que teóricamente se sabe, tiene una fuerte relación con la variable que se mide, requiere identificar un

criterio y la manera adecuada de medirlo, es, obtener las medidas de la variable de interés con el cuestionario y el de la variable criterio para finalmente determinar el grado de relación entre las medidas obtenidas para ambas variables. Esta relación se obtiene mediante el cálculo de un coeficiente de correlación adecuado a las escalas en que se han realizado las mediciones. Mientras mayor sea su valor, mejor será el argumento de validez.

Evidencias basadas en los procesos de respuesta a los ítems

Las evidencias se obtienen analizando los procesos de respuesta que los examinados realizan para contestar a la pregunta. Básicamente se utiliza este tipo de argumento explicativo de la validez en pruebas cognitivas o de habilidad. Para recolectar estas evidencias se requiere de un modelo explicativo de dichos procesos de respuesta, el cual debe guiar todo el proceso de construcción del instrumento de medida. El procedimiento a seguir precisa como punto fundamental, especificar los objetivos de la medición a partir del cual se establecerá un modelo de procesamiento que permitirá generar los ítems o preguntas y realizar un estudio del tipo protocolo verbal en el cual el examinado vaya diciendo en voz alta todos los pasos seguidos para contestar la pregunta o mediante un estudio comparativo del grado de ajuste entre un modelo teórico simulado y el modelo empírico.

Evidencias basadas en las consecuencias de la aplicación del instrumento

La evaluación de las consecuencias del uso de los tests en los procesos de validación es uno de los mayores desafíos planteados por la versión actual de la teoría de la validez y las normas establecidas en los *Standards for educational and psychological testing*(AERA, APA y NCME; 1999). es un desafío conceptual y metodológico, al situar la construcción y uso de los tests en un escenario donde resulta difícil diferenciar entre las cuestiones de validez y los argumentos ideológicos, políticos o sobre la justicia en el uso de los tests, razón por la cual no deja aun de causar desencuentros en las posturas de los especialistas en el campo de la medición.

Aun cuando la validez consecuencial aparece en los estándares del año 1999 es muy poco el avance en cuanto a procedimientos que permitan poner en práctica la recogida de este tipo de evidencias. El trabajo de Padilla, Gómez, Hidalgo y Muñiz (2007) representa un aporte notable en esta dirección al plantear una propuesta metodológica para la evaluación de las consecuencias. Afirman que la distinción entre «inferencias semánticas» e «inferencias políticas» permite integrar la validación de las consecuencias en un esquema único de validación. El proceso de validación debe aportar evidencias sobre los supuestos que sostienen ambos tipos de inferencias de modo que del cuestionario, su forma de valorar las puntuaciones y de analizarlas deben ser consensuadas tanto por el constructor y aplicador como los organismos involucrados en las decisiones que deban tomarse basados en las puntuaciones obtenidas con el instrumento.

LA CONFIABILIDAD DE LAS PUNTUACIONES

Cuando se realiza una medición, ya sea de una variable de tipo físico, biológico psicológico o social se obtiene como resultado un valor que refleja el nivel del sujeto medido en esa variable. No obstante, el procedimiento conlleva inevitablemente un margen de error, un desvío de la precisión. La confiabilidad está asociada a la precisión de la medida obtenida. La precisión es una característica buscada en todo instrumento de medición, sea un termómetro, un reloj o un test. Mientras mayor sea el margen de error de la medida tendremos menos confianza en ella, será una medida inestable, lo cual significa que al repetir el proceso dará valores diferentes y tendremos dudas sobre cual valor asumir como cierto. Por el contrario, un instrumento de precisión con el que se obtengan medidas con márgenes de error mínimos resulta fiable y es lo que aspiramos de un buen test. El objetivo central de los estudios de confiabilidad es la estimación de los errores aleatorios cometidos al medir las variables de interés, estimación que se realiza de distinta manera según la teoría de los tests que se emplee para su estudio.

Confiabilidad desde la perspectiva de la Teoría Clásica de los Tests

Los nuevos trabajos sobre confiabilidad desde la perspectiva de la teoría clásica de los tests muestran que lo que ha habido, sobre todo, es extensiones y refinamientos sobre el tema, más que nuevos desarrollos. Estos refinamientos tienen que ver con el gran poder y bajo costo computacional que han permitido los avances tecnológicos con el uso de computadoras y que han permitido facilitar los cálculos independientemente de las fuentes de error consideradas (Haertel, 2006). En la teoría clásica de los tests (TCT), la más utilizada en nuestro país, se ha prestado atención básicamente a dos fuentes de error:el error debido a las fluctuaciones aleatorias que ocurren a lo largo del tiempo y el error debido a las diferencias en contenido entre los ítems del test.

Error debido a las fluctuaciones aleatorias que ocurren a lo largo del tiempo.

Desde esta perspectiva se ha conceptualizado la confiabilidad como estabilidad temporal, como coherencia de la conducta de los sujetos en distintas mediciones, como consistencia de la escala a lo largo del tiempo.

Los procedimientos propuestos para estimarla han sido el test-retest y el método de las formas paralelas. En el primero se pasa el cuestionario en dos oportunidades y se calcula el coeficiente de correlación entre las puntuaciones obtenidas en cada ocasión. En el segundo se pasan formas que se suponen son equivalentes, aunque los ítems sean diferentes, e igualmente se calcula el coeficiente de correlación. El valor del coeficiente de correlación que se obtenga por cualquiera de los dos procedimientos debe ser suficientemente alto para concluir que las puntuaciones obtenidas son estables, en consecuencia, son confiables.

Error debido a las diferencias en contenido entre los ítems del test.

Desde esta perspectiva se ha conceptualizado la confiabilidad como consistencia interna, como congruencia entre las respuestas de cada sujeto a los distintos ítems del cuestionario.

Los procedimientos propuestos para estimarla se pueden agrupar en dos grandes bloques, los basados en la división del test y los basados en la covarianza de los ítems.

Cuando se utilizan métodos basados en la división en dos mitades, el test se aplica una sola vez, estimándose la fiabilidad al correlacionar las puntuaciones obtenidas por los sujetos en cada una de las dos mitades que conforman el cuestionario. Dado que el test tiene muchas posibles mitades, lo que se acostumbra es dividir el test en ítems pares e impares y la correlación obtenida se corrige mediante la fórmula de Spearman-Brown y se obtiene el coeficiente de confiabilidad.

Dentro de los métodos basados en las covarianzas entre los ítems se incluye el coeficiente alfa formulado por Lee Cronbach en 1951, las ecuaciones formuladas por Kuder y Richardson conocidas como KR20 y KR21, los coeficientes lambda de Guttman, los coeficientes basados en el análisis factorial y la estimación del coeficiente de confiabilidad obtenida a partir del análisis de varianza. De todos ellos, el coeficiente estrella es el coeficiente alfa. No obstante, estos índices deben ser utilizados atendiendo a la escala de respuesta utilizada y sobre todo al hecho de que estos índices trabajan bajo el supuesto de que las escalas son unidimensionales. Si en un test se miden variadas dimensiones del constructo o variable es necesario calcular el índice dividiendo el test en las diferentes escalas que constituye cada dimensión. Varios de estos índices pueden obtenerse directamente con el software SPSS en cualquiera de sus versiones.

Factores de los que depende la confiabilidad estimada desde la TCT

El valor estimado de los índices de confiabilidad depende de la longitud del test, de la muestra de ítems utilizados y de la muestra de sujetos examinados. Dado que depende de la longitud del test, un test con muy pocos ítems podría dar medidas imprecisas, lo cual no siempre es necesariamente así, pero es una cuestión a considerar al momento de elaborar los ítems de una dimensión. No obstante, alargar el tests innecesariamente puede ser visto como una actitud poco ética que sólo busca incrementar artificialmente la confiabilidad. Dado que depende de

los ítems, debe estimarse un coeficiente para cada grupo de ítems que conformen una dimensión, no para la prueba en conjunto, a menos que esta sea unidimensional. Por cuanto depende de los sujetos examinados, debe estimarse el coeficiente con todos los sujetos que conforman el estudio, puesto que grupos diferentes de sujetos darán diferentes valores para el coeficiente. No debe el investigador reportar el valor de la prueba piloto (si la hiciere) sino el de la muestra definitiva.

Confiabilidad desde la perspectiva de la Teoría de Respuesta al Ítem

La Teoría de Respuesta al Ítem (TRI) utiliza un enfoque distinto en la forma de cuantificar el error de medición. Se estima el valor del error para cada sujeto en cada nivel posible de la variable medida y se expresa mediante la inversa de la función de información del test. Desde la perspectiva de la TRI, cuanta más información proporcione el test sobre el atributo medido más fiable puede ser considerado.La función de información y el error estándar en la TRI describen la precisión de las puntuaciones obtenidas con el test como una característica que, a diferencia de lo que sucede con la TCT, varía a lo largo del continuo del rasgo latente y no como un valor único que caracteriza un instrumento de medida, dotando el concepto de confiabilidad de flexibilidad al permitir establecer la precisión que obtendremos en cada nivel del rasgo latente donde se ubique el sujeto.

Cabe decir también que el planteamiento de la teoría de respuesta al ítem ha permitido importantes avances técnicos en el campo de la construcción de tests, como es el caso del establecimiento de bancos de ítems que posibilitan el uso de tests adaptados al nivel del examinado en el atributo que se mide.

CONSIDERACIONES FINALES

La validez y la confiabilidad se estiman para aquellas puntuaciones obtenidas de instrumentos para caracterizar personas, pues la evaluación directa de variables como conocimientos, habilidades, percepciones u opiniones no es posible a menos que se utilicen los instru-

mentos o tests como herramientas de medición. En tal sentido, se recalca la importancia de que la validez de las inferencias y la confiabilidad de las medidas sean analizadas dentro del contexto de los fines para los que se hace la medición, concebidas en relación al contexto y los propósitos de aplicación del test.

Es imperativo insistir en la necesidad de utilizar los términos adecuadamente, tanto desde el punto de vista conceptual, como desde el metodológico. Incurrir en fallas de esta naturaleza podría llevar a que las pruebas reunidas en torno a la validez se centren en lo relacionado con el contenido del test, ignorando otras fuentes potenciales deinvalidez. Asimismo, el no considerar el uso que se les dará a las puntuaciones podría ocasionar que se usen instrumentos que resultaron con una alta precisión en determinadas condiciones para tomar decisiones en otro contexto, sin tomar en cuenta si su uso es apropiado o no en esas nuevas condiciones, en las cuales la confiabilidad podría quedar seriamente afectada.

Es muy importante atender a los criterios de calidad métrica que garanticen la validez de las interpretaciones y la confiabilidad de las medidas (no del instrumento), de lo contrario la calidad de la información en que se sustentan los hallazgos y conclusiones de una investigación generará dudas, propiciando que la credibilidad de la investigación quede en tela de juicio. En la medida que el investigador aporte la mayor cantidad de evidencias que argumenten la validez de las interpretaciones y maximice la precisión de la medida, mayor solidez tendrán las conclusiones emanadas de su estudio.

REFERENCIAS BIBLIOGRÁFICAS

Abad, J., Olea, J., Ponsoda, V. y García, C. (2011). *Medición en ciencias sociales y de la salud.* Madrid: Editorial Síntesis.

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

American Psychological Association.(1954). *Technical recommendations for psychological test and diagnostic techniques*. Washington, DC: Autor.

Brennan, R. L. (Ed.) (2006). *Educational measurement* (4ª ed.). Washington, DC: National Council on Measurement in Education and American.

Frisbie, D. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice, 24*, 21 -28.

Haertel, E. H. (2006). Reliability. En: R. L. Brennan (Ed.), *Educational measurement* (4ª ed., pp. 65–110). Washington, DC: National Council on Measurement in Education and American.

Kane, M. (2006).Validation. En: R. L. Brennan (Ed.), *Educational measurement* (4ª ed., pp. 17-64). Washington, DC: National Council on Measurement in Education and American.

Linn, R. L. (Ed.) (1989). *Educational measurement (3ª ed.)*. New York: Macmillan.

Messick, S. (1989).Validity. En R.L. Linn (Ed.), *Educational measurement* (3ª ed., pp.13-103). New York: Macmillan.

Padilla, J. L., Gómez, J., Hidalgo, M. D. y Muñiz, J. (2007). Esquema conceptual y procedimientos para analizar la validez de las consecuencias del uso de los tests. *Psicothema, 19,* 173 -178