Pag. 15-33 4/27/06 12:20 PM Page 15 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

SUBJETIVIDAD Y PROCESOS COGNITIVOS, 2006
Pag. 15-33
UCES

TECNICAS PSICOMETRICAS. CUESTIONES DE VALIDEZ Y CONFIABILIDAD

Juan Carlos Argibay*

Resumen

En este artículo se analizan cuestiones de validez y confiabilidad referidas a las pruebas psicométricas. Se comentan los distintos aspectos de la confiabilidad, ya sea en lo que hace a su consistencia interna, como a la estabilidad de las mismas y a cuestiones de equivalencia. En cada caso se discuten sus características generales y se plantean, brevemente algunos de los distintos procedimientos para estimar la con - fiabilidad según de qué tipo se trate. En cuanto a la validez se indaga en sus princi - pales tipos: de contenido, de criterio y de constructo y como se vinculan entre sí. Se plantean también los nexos entre validez y confiabilidad. Tratamos de plantear co - mo estos estudios se vincularían con una visión metodológica global, en cuanto a métodos de investigación y diseños y respecto de su funcionalidad como formas de operacionalizar constructos hipotéticos. Dentro de la validez de criterio, se analiza tanto la concurrente como la predictiva y se las relaciona con conceptos como el po - der predictivo y poder explicativo. La validez de constructo es estudiada en los pro - blemas que presenta y en las distintas formas para establecer la misma, consideran - do la validez relacionada con la estructura factorial y la validez convergente y di - vergente.

Palabras clave: *técnicas psicométricas; validez; confiabilidad; validez de construc - to; validez de contenido; validez de criterio.*

Abstract

In this article questions of validity and reliability referred to the psychometric´s tests are analysed. Are commented the different aspects of the reliability, whether in which does to their internal consistency, as to the stability of the same and to questions of equivalence. In each case its generals characteristics are discussed and they are briefly presented some of the different procedures to reckon the reliability depending on the type. As for the validity is investigated in their main types: of content, of cri - terion and of construct and as they are linked among if. Apart from that, links bet -

* UBA– UCES. E-mail: jcargibay@hotmail.com

Pag. 15-33 4/27/06 12:20 PM Page 16 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

ween validity and reliability are presented. We try to present how these studies can be links with a methodological global´s vision, as for methods of investigation and design and as for its functionality as a way of operationalizer hypothetic´s cons - tructs. Inside the opinion of validity is analysed so much the concurrent as the pre - dictive and relates itself it to concepts as the predict and explanation´s power. The construct validity is studied in the problems that presents and in the different forms to establish the same one, considering the validity related to the factorial´s structu - re and the convergent and divergent validity.

Key words: *psychometric´s tests; validity; reliability; construct validity; content va - lidity; criterion validity.*

Al aplicar una técnica psicométrica, independientemente del propósito con que se utilice, se juegan permanentemente, cuestiones relacionadas con la validez y confiabilidad del instrumento.

Es cierto que muchas veces se utiliza un test, sin tener mucho en cuenta o, analizar su validez y confiabilidad. Pero se debiera considerar que las bondades de aquél, dependen mucho de estos factores. Ya sea que se utilice la técnica con fines puramente psicodiagnósticos o en investigación.

Para comenzar, podríamos preguntarnos, ¿cuál es el propósito de una técnica de evaluación, ya sea psicométrica o proyectiva? Consideramos que, en principio, una respuesta adecuada sería: la técnica sirve para operacionalizar algún constructo hipotético. Por ej., al evaluar la variable de personalidad Extraversión, mediante alguna técnica psicométrica, lo que estamos haciendo es operacionalizar mediante un instrumento dicho constructo hipotético. Puede haber más de una forma de operacionalizar dicho constructo y, de hecho, tanto el EPQ-Acomo el Neo-Pi-R miden entre otras dicha variable. Cabría preguntarse si, en realidad, miden el mismo constructo. Volveremos sobre este tema al analizar más adelante la validez de constructo.

Considerando lo anterior, es de fundamental importancia que el instrumento esté midiendo lo que dice medir, caso contrario se estaría operacionalizando incorrectamente el atributo, al no corresponderse lo que realmente se observa, con lo que se cree medir. También podríamos preguntarnos, ya que toda medida tiende a tener errores de medición, en qué medida el instrumento que estamos usando es preciso en medir los valores verdaderos de la variable que se analiza y cual es su grado de congruencia para medir la variable en cuestión.

Esto último, la precisión y la congruencia, tienen que ver con la confiabilidad del ins-

Pag. 15-33 4/27/06 12:20 PM Page 17 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

trumento. Que el instrumento mida realmente lo que dice medir, se relaciona con la validez.

Existen distintas formas de validez y confiabilidad. A continuación pasaremos a analizarlas.

Confiabilidad

Como ya mencionamos, toda medición tiende a presentar errores, de manera que el valor observado en la medición, está formado por el valor verdadero y por el error de medición. Si se pudiera llegar a una medición que, en un sentido ideal no tuviera márgenes de error, el valor observado y el valor verdadero (que se correspondería con la variable a medir), coincidirían. Pero esto es algo ideal, ya que las medidas incluyen error. Obviamente, cuanto más error incluya la medición, más contaminada quedaría la variable de interés y más imprecisa sería la técnica que intenta medirla. Por éj.: si aplicamos un test que mide la variable de personalidad Neuroticismo y obtenemos un puntaje de 15 (valor observado) y el valor real de la persona fuera de 14 puntos (valor verdadero), tendríamos un margen de error de 1 punto; en otro caso, si el valor real fuera también de 14 puntos y hubiéramos obtenido en la aplicación del instrumento un puntaje de 10, el error de medida sería de 4 puntos. De manera que al aumentar el error de medida, el puntaje que obtenemos al aplicar el test se encuentra más alejado de la medida que sería la verdadera, o sea, que sería más impreciso.

Podemos entonces vincular la confiabilidad con los errores de medición de la siguiente manera: si en toda medida el valor obtenido, está compuesto por el valor verdadero y los errores de medición, un instrumento será más confiable, en la medida que maximice el valor verdadero. Con maximizar el valor verdadero, nos referimos a que éste aproxime lo más posible al valor observado u obtenido. Al reducir el error de medición, incrementamos la similitud entre el valor verdadero y el obtenido. De manera que al maximizar el valor verdadero, minimizamos el error de medición, con el consecuente incremento de la confiabilidad.

Es importante tener en cuenta, que la confiabilidad se refiere, específicamente a los ***errores aleatorios***, tal cual menciona Martínez Arias (1996). Podemos hablar de dos tipos de errores: los errores aleatorios que, como tales, no pueden ser controlados y no se pueden predecir y los errores sistemáticos que son controlables y pueden ser explicados por alguna fuente de variación sistemática. De ambos errores los únicos que interesan a la teoría de la fiabilidad son los errores aleatorios.

Para Martínez Arias (1996), la confiabilidad es ***la consistencia en un conjunto de medidas de un atributo***. Podríamos también definir la confiabilidad como la propor-

Pag. 15-33 4/27/06 12:20 PM Page 18 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

ción de la variabilidad verdadera respecto de la variabilidad obtenida.
En el análisis de la confiabilidad tenemos que considerar tres aspectos:

- ***Su congruencia o consistencia interna.***
- ***La estabilidad del instrumento.***
- ***La equivalencia***

Congruencia o consistencia interna

La congruencia interna consiste en que las distintas partes que componen el instrumento estén midiendo lo mismo. Por éj., en un test compuesto por distintos reactivos, se esperaría que cada uno de ellos midiera el mismo atributo. En un instrumento que incluyera distintos factores, se esperaría que aquellos reactivos que remiten al mismo factor estuvieran midiendo también lo mismo. La forma de pensar la congruencia interna, se basa en que las distintas partes del instrumento correlacionen entre sí. Al hablar de confiabilidad del instrumento, estamos hablando de que los reactivos que componen el instrumento, son distintas formas del mismo atributo, de manera que estaríamos frente al mismo constructo, observado en conductas diferentes. De manera que en condiciones ideales, esperaríamos encontrar que los distintos ítems estuvieran correlacionados entre sí. Podemos concluir, entonces, que la congruencia interna del instrumento se establecería según la magnitud de los valores que expresaren las correlaciones entre sus partes.

Por ejemplo: en el test EPQ-A de Eysenck (TEA Ediciones, 1982) las preguntas: ¿Tiene Ud. muchos "hobbys", muchas aficiones?, ¿Es Ud. una persona conversadora?, ¿Es Ud. una persona animada, alegre?, remiten junto con otras al factor Extraversión, en consecuencia, se esperaría que las respuestas que dieran los sujetos a estas preguntas estuvieran relacionadas entre sí. Si las preguntas no correlacionaran eso afectaría la confiabilidad del instrumento.

Un comentario respecto de lo anterior: en la teoría clásica de la fiabilidad se parte de una situación ideal en la que todos los ítems del instrumento no solo medirían lo mismo sino que lo harían con idéntica intensidad. Lo que ocurre realmente es que, más allá de que estén midiendo lo mismo, la intensidad con que lo hacen varia de ítem a ítem y, además, es usual que midan más de un constructo o dimensión, como menciona Morales (1988). Lo anterior también influye en una reducción de los niveles de correlación y consecuentemente en una reducción de la confiabilidad.

La confiabilidad a través de la congruencia interna, es una de las más utilizadas por los investigadores, entre otras cosas, por requerir una sola aplicación del instrumento, lo cual facilita la investigación.

Pag. 15-33 4/27/06 12:20 PM Page 19 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

Los procedimientos para calcular la congruencia interna podemos dividirlos en dos métodos principales:

- Métodos basados en la división del instrumento en dos mitades.
- Métodos basados en la covarianza de los ítems.

El método de la división en mitades consiste en lo siguiente:

1º. Se divide el instrumento en dos mitades.

2º. Se obtiene la puntuación para cada mitad en forma independiente.

3º. Se usan estos puntajes para estimar el coeficiente de correlación entre ambas mitades. Este coeficiente de correlación se ajusta mediante la fórmula de Spearman-Brown.

El motivo por el cual es necesario aplicar la fórmula de Spearman-Brown es que "*las escalas de mayor longitud son más confiables que las cortas"* (Polit y Hungler, 1995/1997); por ese motivo el coeficiente de correlación que fue obtenido a partir de las dos mitades, tiende a subestimar el valor real de la confiabilidad, que se obtendría con una escala con el doble de longitud que la de cada una de las mitades y que constituiría la longitud original del instrumento.

Existen distintas formas de dividir el test en dos mitades, y cada una de ellas nos da un valor diferente del coeficiente de confiabilidad. Si bien hay que aclarar que si la confiabilidad es buena, las variaciones del coeficiente en función de la forma de dividir el test serían reducidas. De todos modos, en función de lo anterior, se desarrollaron los métodos basados en la covarianza de los ítems, que solucionan el problema anterior al dividir por cada ítem, lo que permite, tal cual dice Martínez Arias (1996), "*tratar a cada ítem como si fuese un test de longitud unidad con una puntuación."*

Dentro de los métodos basados en la covarianza de los ítems el más utilizado es el coeficiente alpha de Cronbach y en general es preferible al método de la división por mitades. Si bien, de ser confiable el instrumento, con ambos métodos se obtendría un coeficiente de confiabilidad aceptable.

Algo que es fundamental al interpretar el coeficiente alpha es que el valor de éste, se ve afectado por el número de ítems. De manera que entre dos tests que tuvieran realmente la misma confiabilidad, el que tuviera menos ítems daría un menor coeficien-

Pag. 15-33 4/27/06 12:20 PM Page 20 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

te alpha. Esto se debe a que al aumentar la longitud del instrumento la varianza verdadera aumenta en mayor proporción que la varianza de error. Debido a lo anterior, Cronbach propuso una fórmula adicional de consistencia inter-ítem, que consistiría en una estimación de la correlación media inter-ítem y no se vería afectada por la longitud del instrumento. Esta nueva fórmula no llegó a alcanzar la difusión del coeficiente alpha, pero consideramos que puede ser útil su aplicación, cuando se desea comparar la confiabilidad de dos instrumentos con número desigual de ítems, o de dos factores dentro de un instrumento con un número de reactivos desigual.

En cuanto a cuál debe ser el valor mínimo del coeficiente alpha, no hay un criterio uniforme. Podríamos decir que un parámetro aceptable es un mínimo de .70, pero esto hay que tomarlo con precaución, ya que habría que considerar el número de ítems del instrumento. Si su longitud es reducida, valores menores a .70, podrían estar indicando una buena confiabilidad y si su longitud fuera extensa, un valor de .70 podría no ser tan buen indicador de confiabilidad. Por ejemplo si tuviéramos un instrumento con 10 ítems y alpha = .60 y otro con 100 ítems y alpha = .80 y calculáramos para cada uno la correlación media inter-ítem. Esta sería para el primero de .1304 y para el segundo de .0385; de manera que, si bien el segundo tiene un coeficiente de confiabilidad mayor, atendiendo a la cantidad de ítems, en realidad es menos confiable que el primero, si consideramos la correlación media inter-ítem. Además se debería ser más exigente respecto a la confiabilidad, si la prueba va a ser utilizada para tomar decisiones con fines psicodiagnósticos individuales, que si va ser utilizada grupalmente, para obtener con la misma, valores medios grupales.

De manera que, posiblemente, lo más recomendable al interpretar el coeficiente de confiabilidad sea considerar cuan próximo está al valor máximo (uno), evaluar el margen de error y considerar las distintas circunstancias que involucren el uso del instrumento y, fundamentalmente, tener en cuenta el número de ítems que contenga el instrumento.

En la práctica, es muy importante que los instrumentos sean confiables. Los tests son utilizados en diversas áreas y, en todos los casos, quien los utiliza supone que el mismo mide con precisión determinado atributo, éste es el sentido de su aplicación. Pero, si el instrumento es poco confiable y entonces poco preciso y se lo usara por éj., para una investigación, podría ocurrir, de no cumplirse las hipótesis planteadas, que esto se debiera no a defecto de las mismas, sino a error de los instrumentos de medición, por ser poco confiables. Además, si hubiera que tomar decisiones respecto de un caso individual y, como parte de los elementos de juicio se utilizara un test, el riesgo de tomar una decisión equivocada, se vería incrementado por la poca confiabilidad del instrumento. Además, la confiabilidad y validez tienen cierta relación entre

Pag. 15-33 4/27/06 12:20 PM Page 21 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

sí. De manera que si un instrumento es poco confiable, esto podría afectar su validez. Si un instrumento es poco confiable está midiendo, aparte del atributo que pretende medir, diversas variables que son fuente de errores aleatorios, lo cual distorsionaría la medición de la variable de interés. Ya que como sugieren Polit y Hungler (1995/1997) si el instrumento es errático, inconsistente e impreciso difícilmente pueda medir con validez el atributo en cuestión. Esto no quiere decir que si un instrumento tiene alta confiabilidad, eso signifique que tenga que ser válido; un instrumento puede ser confiable y al mismo tiempo no ser válido. Pero a la baja confiabilidad, se la puede relacionar con problemas de validez y, justamente, si la aplicación de un instrumento es para operacionalizar un constructo, el que tenga baja confiabilidad debe servir como un alerta respecto de la validez del mismo y a la conveniencia de aplicar el instrumento para medir el correspondiente atributo.

Estabilidad

Al analizarse la congruencia interna, en ningún momento se contemplaban las variaciones que podían darse en las medidas y la consecuente disminución de la confiabilidad, debidas al paso del tiempo. Al hablar de la estabilidad de un instrumento contemplamos este factor, y lo que se observa es en qué grado se obtienen las mismas medidas al aplicar dos veces el mismo instrumento, mediando entre ambas tomas un tiempo determinado. Entre ambas tomas pueden intervenir una serie de factores que sean fuentes de varianza de error, disminuyendo la confiabilidad de las medidas.

El procedimiento es relativamente sencillo, consiste en aplicar la misma prueba en dos momentos diferentes al mismo grupo de sujetos y luego, correlacionar entre sí, los puntajes obtenidos en cada toma. Esta correlación sirve para establecer la confiabilidad del instrumento.

En este caso no estamos hablando solamente de que las partes del test correlacionen en forma significativa. Ya que no se trata de comprobar la relación entre dos variables diferentes, en cuyo caso si bien sería importante la magnitud de la correlación, importaría que se pudiera comprobar que hubiera correlación. Además de que si fueran dos variables diferentes, no se esperaría necesariamente que la correlación fuera alta, ya que estarían presentes otras fuentes de variación, aparte de la varianza de error. En este caso es la aplicación del mismo test, o sea que las variables a relacionar son las mismas. De manera que en condiciones ideales, esperaríamos encontrar una correlación perfecta. Esto no ocurre, porque debido a los errores de medición (anteriormente mencionados), hay un porcentaje de error aleatorio, incluido en la medición. No obstante, la correlación debiera ser en estos casos alta, ya que se esperaría que estas fuentes de error se vieran minimizadas, y que en la aplicación del instrumento no se incurriera en errores sistemáticos.

Pag. 15-33 4/27/06 12:20 PM Page 22 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

Hay varias cuestiones que podrían afectar la confiabilidad en estos estudios test-retest:

- Modificaciones en la actitud de los sujetos respecto de la prueba o cambios en la información que éstos manejen podrían afectar los puntajes de la segunda toma del test, incidiendo sobre los valores de correlación, ya sea incrementando artificialmente la confiabilidad o infravalorándola.

- Puede haber efectos derivados de la primera aplicación del test que afecten el rendimiento en el segundo. Por. éj., que algunos sujetos recuerden las respuestas y al contestar el segundo test tiendan a poner las mismas respuestas, no tanto por que se correspondan con su perfil en el atributo que se mide, sino por contestarlo igual. También en tests que midan aptitudes o destrezas los sujetos pueden ver condicionado su rendimiento por la práctica previa.

- Un tema de fundamental importancia es el lapso que medie entre ambas tomas. Un lapso demasiado breve podría incrementar el efecto distorsivo de la memoria y la práctica. Si bien esto se solucionaría incrementando el tiempo, esto a su vez produciría otro problema, la aparición de variables relacionadas con la maduración y la historia:

  - La maduración que, según define Arnau Gras (1982*), "abarca el conjunto de procesos biológicos y psicológicos que se operan en los sujetos como conse - cuencia del paso del tiempo."* Suele ser mucho más relevante si los sujetos son niños, ya que en ellos se pueden esperar cambios pronunciados en lapsos de tiempo breve.

  - La historia incluye todos aquellos eventos que pudieran tener lugar en el lapso que medie entre ambas aplicaciones y que podrían llegar a afectar el resultado de la segunda prueba. Esto se manifestaría menos en atributos estables como rasgos de personalidad, que en aquellos que describieran estados. Por éj., si consideramos la variable ansiedad, la forma en que podría intervenir la historia produciendo modificaciones en los puntajes de la segunda toma, sería muy diferente si se tratara de un test que midiera la ansiedad como un rasgo de personalidad, en cuyo caso se esperaría que éste fuera estable y que no sufriera cambios por el paso del tiempo o que fueran ínfimos, que si se evaluara la ansiedad como estado, la cual puede sufrir cambios importantes con el transcurso del tiempo.

  - En el caso de instrumentos que evaluaran psicopatologías, la historia podría afectar la confiabilidad test-retest, principalmente en aquellas psicopatologías

Pag. 15-33 4/27/06 12:20 PM Page 23 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

(de carácter neurótico), con altos índices de remisión espontánea, no tanto en aquellas con remisión espontánea baja o casi nula. Por ej, Eysenck (1977/1978) menciona que las fobias específicas y los trastornos obsesivo-compulsivos, muestran una recuperación particularmente lenta, y posiblemente muy a menudo no haya ninguna recuperación sin tratamiento.

Uno de los problemas de las técnicas para evaluar la confiabilidad mediante test-retest es que pueden llegar a confundir las fluctuaciones aleatorias con cambios que se producen realmente en el atributo. Los cambios que se podrían producir en el atributo, modificarían (en aquellos sujetos en que se produjeran), los puntajes en el instrumento que se estuviera evaluando, con la consecuente disminución de la confiabilidad. Pero en este caso la disminución de la confiabilidad sería ficticia, no se correspondería con errores aleatorios, sino con errores sistemáticos, los cuales no son de interés de la teoría de la fiabilidad. En este caso hablamos de errores sistemáticos desde el punto de vista de la confiabilidad, pero en otro sentido, representarían realmente varianza sistemática producida en la variable que evalúe el test.

Por ejemplo: supongamos que se evaluara la confiabilidad de un test que midiera estado depresivo y algunos de los sujetos estuvieran en tratamiento por dicha psicopatología, o por otra. Ya sea en un caso directamente, o en el otro indirectamente, la intervención podría producir una modificación en el atributo (estado depresivo) y esto reduciría los valores de confiabilidad. Pero la disminución de los puntajes de depresión sería esperable como un efecto de la intervención terapéutica y de ninguna manera podrían tomarse como un indicador real de problemas de confiabilidad del instrumento. También podría ocurrir que el instrumento fuera utilizado para hacer un seguimiento de la evolución de pacientes o en investigaciones sobre eficacia terapéutica. En estos casos se esperaría una disminución de los puntajes como efecto del tratamiento e, idealmente, en la mayoría de los sujetos. Estos tampoco serían errores aleatorios, sino, sistemáticos y no serían representativos de la confiabilidad del instrumento. Cabe aclarar, en este último caso, que no toda variación sistemática, tendría que afectar necesariamente los valores de confiabilidad que se obtendrían sin esta variación. Si el efecto del tratamiento se cumpliera de manera que produjera una mejora en todos los sujetos y en el mismo grado, los valores de confiabilidad no se verían afectados por el tratamiento, ya que sería la misma que si este no hubiera tenido lugar, debido a que al ser la disminución idéntica para todos (en un valor constante), por éj., la puntuación del test disminuyera en todos los sujetos diez puntos, la correlación entre ambas pruebas sería la misma, que la que se hubiera obtenido si el tratamiento no hubiera tenido lugar. Lo que realmente afectaría la confiabilidad sería el efecto diferencial del tratamiento, ya que con el mismo, algunos tendrían una mejoría mayor que otros, algunos ninguna mejoría y algunos hasta podrían empeorar;

Pag. 15-33 4/27/06 12:20 PM Page 24 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

este efecto dispar del tratamiento, alteraría el ordenamiento que los sujetos hubieran tenido sin él, y modificaría los valores de correlación. Esto último es lo que ocurre en la realidad, el primer caso que implica una modificación igual en todos los sujetos, es un caso ideal que no se corresponde con lo que realmente ocurre en un tratamiento. Por tal motivo la aplicación de un tratamiento psicológico en algunos de los sujetos, entre las dos tomas de una investigación test-retest, de un instrumento que midiera psicopatología, podría afectar los valores de confiabilidad que se obtuvieran para dicho instrumento.

Lo anterior, podría hacerse claramente extensivo, a las diversas técnicas psicométricas que miden distintas psicopatologías. Además es importante en estas pruebas conocer su estabilidad pues, de ser usadas, por éj., para ir evaluando un tratamiento a través del tiempo, sería fundamental saber cuanto afecta el paso del tiempo a la confiabilidad del mismo, principalmente cuando se use en la clínica, ya que en investigación el disponer de grupo control, permitiría un mejor manejo de ese problema.

Puede haber también otras fuentes de errores sistemáticos, que alteren los puntajes del test, sin modificar básicamente el atributo, como ser por éj., factores que modifiquen el contexto de la medida. Supongamos el siguiente ejemplo: un psicólogo aplica un test que mide Atención a un grupo de sujetos. Dos días más tarde les aplica de nuevo el mismo test, pero cambia el lugar donde lo hace, y en este lugar hay mayor cantidad de ruido ambiente. Es probable que el resultado en la prueba de atención sea mejor en la primera toma, ya que los sujetos estaban expuestos en menor medida a estímulos distractores que pudieran afectar su rendimiento.

En los ejemplos anteriores, podemos observar, como al introducir una variable en forma sistemática entre ambas aplicaciones del test, se podrían alterar los resultados obtenidos en la confiabilidad test-retest. En qué casos los efectos de la historia producirían una reducción ficticia de los valores de confiabilidad y en qué casos estarían afectando realmente la confiabilidad. La diferencia que podemos marcar es la siguiente, en los ejemplos que dimos, la modificación en el contexto de medida y la intervención terapéutica, estamos frente a variables, a la que están expuestos todos los sujetos o una parte de ellos, y que producen un cambio sistemático, consistente en una reducción de los puntajes (en otros ejemplos podría haber sido un incremento), como una tendencia grupal; ya que como mencionamos con anterioridad tomando cada caso en particular el efecto podría ser diferente en cada sujeto, fundamentalmente en grado (en qué medida se produce la modificación) y en una minoría de casos en dirección (por éj., si la mayoría experimenta una reducción de los puntajes, algunos podrían mantener el mismo puntaje o incrementarlo).

Pag. 15-33 4/27/06 12:20 PM Page 25 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

Por el contrario, cuando hablamos de la historia, nos referimos a un conjunto amplio de variables, donde cada sujeto se ve expuesto a distintas variables y/o a combinaciones diferentes de las mismas, sin poder determinarse cual o cuales variables afectaron a cada uno y de que forma. Se podría producir un incremento en el error, pero sin poder determinar una fuente de error claramente identificable, o sea, que los cambios en los puntajes serían impredecibles e incontrolables y, en consecuencia, errores aleatorios.

Consideramos entonces, que los efectos no identificados derivados de la historia y que contribuyeran a la varianza de error, de producir un descenso en la correlación entre ambas tomas, constituirían un déficit real de confiabilidad del instrumento, que deberá ser contemplado, por ejemplo, cuando el instrumento se utilice para evaluar el transcurso de un tratamiento, como así también en otras aplicaciones del test. Por el contrario, las diferencias en las medidas, claramente atribuibles, a fuentes de variación sistemática, deberían ser controladas dentro de lo posible, para que no sean incluidas como déficit de confiabilidad.

Equivalencia

La técnica de equivalencia se aplica cuando se quiere determinar la confiabilidad de dos instrumentos que se consideran paralelos. Se trata de poder establecer la consistencia o equivalencia de los instrumentos, que suponen medir el mismo constructo, al aplicarlos a los mismos sujetos, Polit y Hungler (1995/1997). En este caso se aplican las dos formas del test en forma simultánea, se va variando el orden de presentación de las formas de sujeto en sujeto (para controlar posibles efectos de orden) y luego se correlacionan los puntajes de ambas formas. De ser ambas formas equivalentes tendría que obtenerse un coeficiente de correlación elevado. En la práctica es muy difícil establecer que las formas sean realmente paralelas, por tal motivo suele tratárselas más bien como formas alternativas. A diferencia de las otras técnicas de confiabilidad, en ésta, no solo se establece la confiabilidad, respecto de la magnitud de los errores aleatorios, sino que también se establece cual es el grado de paralelismo que hay entre ambas formas.

Aveces ambas formas no se aplican simultáneamente, sino que se deja transcurrir un tiempo entre ambas, de manera que la correlación que se obtendría serviría tanto para estimar la equivalencia como la estabilidad de las dos formas para medir el atributo.

Otra forma de equivalencia consiste en determinar la confiabilidad entre evaluadores u observadores. Esta forma de confiabilidad no se puede aplicar a cualquier instrumento, sino que es específica de aquellos instrumentos en los cuales dos o más evaluadores u observadores tienen que calificar o puntuar la conducta y/o el rendimien-

Pag. 15-33 4/27/06 12:20 PM Page 26 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

to como parte de la aplicación del instrumento. Cabe aclarar que, en este caso, estamos usando el término conducta en un sentido amplio, sin remitir únicamente a la conducta directamente observable. Por ej., si en la aplicación del instrumento, los sujetos se expresaran en un relato sobre ellos mismos, y debiera establecerse una valoración mediante evaluadores, de las conductas descriptas en estos relatos, sería muy importante poder determinar la concordancia entre distintos evaluadores. Hay que tener en cuenta que en estos casos los evaluadores son una importante fuente de error. Para determinar la confiabilidad entre los evaluadores u observadores se usan índices de concordancia, como por ejemplo, el coeficiente Kappa de Cohen.

Validez

Habíamos comentado que los instrumentos tenían como propósito operacionalizar constructos, o sea, medir determinados atributos. La validez tiene que ver con poder determinar si el instrumento está midiendo realmente el atributo que dice medir.

Determinar la validez de un instrumento es mucho más difícil que establecer su confiabilidad. Como dicen Polit y Hungler (1995/1997), *"no es común encontrar pruebas sólidas que sustenten la validez de la mayor parte de las mediciones de carácter psicológico"*. Esto se debe a que mientras la confiabilidad sería esencialmente una cuestión empírica; la validez incluye más elementos teóricos, ya que la validación persigue la explicación, con todas las complicaciones que esto implica.

Dentro de los distintos tipos de validez, encontramos como las más importantes, las siguientes: validez de constructo, validez de criterio y validez de contenido. Es importante tener en cuenta que distintas formas de validez, pueden ser adecuadas para propósitos diferentes y que cada una permite inferencias distintas, que no pueden cambiarse entre sí.

Validez de contenido

Al construir un test, elegimos determinados ítems de un conjunto de conductas que tienen un interés específico, por suponer que remiten al atributo a ser evaluado por el test. En el instrumento no colocamos todas las conductas posibles, elegimos algunas de ellas, o sea, que hacemos una muestra de conductas. Al analizar la validez de contenido, lo que hacemos es evaluar si los ítems que hemos usado para construir el test, son relevantes para el uso que se le va a dar al test, es decir, si todos los ítems están dentro del dominio de interés. Una vez establecida la relevancia, lo que importa es saber si los ítems constituyen una muestra representativa del universo de conductas que podrían haberse elegido como indicadores del atributo en cuestión. La validez de contenido se centra en establecer de la mejor manera lo anterior.

Pag. 15-33 4/27/06 12:20 PM Page 27 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

Validez de criterio

Podemos decir que en el caso de la validez de criterio, lo que se persigue, es un fin más práctico, no se trata únicamente de establecer que se mida adecuadamente un constructo, sino fundamentalmente relacionar las puntuaciones del instrumento con otras variables a las que llamaremos criterio.

La validez de criterio nos es útil especialmente cuando deseamos hacer inferencias a partir de los puntajes que se obtienen en el test respecto de alguna otra variable de interés. Por ejemplo, si nos interesara establecer en que medida los puntajes obtenidos en la variable de personalidad Neuroticismo, nos permiten inferir la predisposición para adquirir determinadas conductas con características psicopatológicas, o en que medida pueden servirnos para predecir una mejor evolución terapéutica, o menores recaídas; relacionaríamos dicha variable de personalidad con los distintos criterios mencionados. Si lo que quisiéramos fuera establecer la validez de criterio de un test de inteligencia, podríamos usar algún indicador de rendimiento académico (por éj. notas escolares), como criterio, para relacionar con el test de inteligencia.

Dentro de la validez de criterio se habla de validez concurrente y validez predictiva. La diferencia entre ambas formas de validez, radica en la temporalidad del criterio. Si las puntuaciones del test se utilizan para predecir alguna medida del criterio que se va a realizar a futuro, sería validez predictiva. Si por el contrario relacionamos las puntuaciones del test con alguna medida del criterio tomada en el mismo momento sería validez concurrente. Por ejemplo, si aplicáramos el EPQ-A (Cuestionario de Personalidad de Eysenck para Adultos), y lo utilizáramos para predecir que pacientes depresivos, pasado un año, tendrán una mejor respuesta al tratamiento, sería validez predictiva. Si por el contrario aplicáramos el EPQ-Ay el Inventario de Depresión de Beck (como criterio), en forma simultánea, y relacionáramos los puntajes de ambos instrumentos entre sí, para analizar en qué medida el Neuroticismo puede predisponer a la adquisición de conductas depresivas, sería validez concurrente.

Nos parece conveniente aclarar lo siguiente: cuando hablamos de validez predictiva, no debemos confundir ésta con el poder predictivo propio de las investigaciones científicas. Si de poder predictivo se trata, éste está presente, tanto en la validez predictiva como en la concurrente. Cuando se ha hecho un estudio de validez concurrente y se ha establecido que tal variable de un test correlaciona con tal criterio, estamos prediciendo esa correlación, y eso es poder predictivo. O sea, que la diferencia entre ambas formas de validez, no tiene que ver con si son predictivas o no (desde un punto de vista científico), sino más bien con el diseño que involucran en cuanto a su dimensión temporal. La validez concurrente implica un diseño transeccional o transversal, los cuales como bien mencionan Sampieri, Collado y Lucio (1997), *"recolec -*

Pag. 15-33 4/27/06 12:20 PM Page 28 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

tan datos en un solo momento, en un tiempo único"; mientras que la validez predictiva implicaría un diseño prospectivo.

Validez de constructo

Gran parte de las variables psicológicas, no son observables directamente, o sea, que constituyen constructos hipotéticos, que forman parte de las diversas teorías que tratan de explicar la conducta humana. Estas variables al no poder ser observadas directamente, para toda investigación, deben ser operacionalizadas, o sea, indicarse los procedimientos de medida para observar la variable, hacerla empírica. Los instrumentos psicométricos se refieren a constructos hipotéticos, siendo el instrumento una forma de operacionalizar los mismos. De esta manera la validez de constructo, consiste en tratar de probar que las conductas que registra el test, pueden ser consideradas indicadores válidos del constructo al cual refieren.

Es importante tener en cuenta que en la validación de constructo, no se trata de hacer corresponder en forma única la puntuación del test con el constructo, ya que un mismo constructo, puede tener, varios indicadores, es decir, varias operacionalizaciones. De lo que se trata es de establecer que las puntuaciones del test constituyen en forma válida una de las manifestaciones del constructo. También hay que tener cuidado, en el sentido de que puede ocurrir, que en algunos casos, diferentes tests que tratan de medir el mismo atributo, no sean en realidad distintas operacionalizaciones del mismo constructo, sino constructos diferentes, lo cual dependerá de las características de la teoría en que cada uno esté inserto, y el grado de similitud entre las mismas.

La validez de constructo, es el principal tipo de validez y a su vez, la más difícil de comprobar. Si queremos establecer que las conductas que registra el test, son indicadores válidos del constructo, tendremos que considerar que sólo podremos aproximar a ese objetivo. Ya que suponer que se puede comprobar que un término conceptual, se corresponde exactamente con un término empírico, implicaría negar la diferencia de niveles entre ambos, que prácticamente imposibilita comprobar una correspondencia exacta. Con esto no queremos decir que ambos términos no puedan corresponderse exactamente, sino que no podríamos justificarlo. Solo aportar una serie de evidencias, más o menos convincentes que apunten en ese sentido.

Además, en este caso, lo que cuenta no son tanto cuestiones de utilidad en la aplicación del instrumento, las cuales eran importantes en la validez de criterio, sino el atributo que subyace a las conductas observables del test. O sea, que la cuestión es eminentemente teórica, y comporta gran parte de los problemas que podemos encontrar en la justificación de teorías.

Pag. 15-33 4/27/06 12:20 PM Page 29 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

Lo que importaría en este caso, si tomamos por ejemplo la variable Neuroticismo medida por el EPQ-A, no sería tanto si las puntuaciones de Neuroticismo pueden servir para establecer predisposición para la adquisición de determinadas psicopatologías. Lo que se consideraría más relevante sería establecer si este factor que mide el test se corresponde con el constructo Neuroticismo, ubicado en la teoría de la Personalidad de Eysenck, de la cual se deriva el instrumento al que nos referimos.

Hay distintos procedimientos para evaluar la validez de constructo, mencionaremos algunos de ellos:

• **Análisis factorial:**
Esta forma de evaluar la validez se utiliza cuando el test está dividido en factores y sirve para medir la validez de constructo, debido a que desde la teoría del instrumento se plantean los distintos factores como atributos diferenciados. Para comprobar la validez de constructo (factorial) se utiliza el Análisis factorial. Esta técnica analiza las intercorrelaciones de un conjunto de datos, para establecer determinadas agrupaciones de ítems correlacionados entre sí, las cuales remiten a factores subyacentes, que no son observables, o sea que constituyen distintos constructos. Estos constructos forman parte de la teoría del test. El análisis factorial, se utiliza no solo para evaluar la validez del instrumento, sino también en su construcción. Una vez construido el test y establecidos los correspondientes factores, la técnica puede aplicarse sobre los datos obtenidos a partir de una muestra de sujetos para establecer si la estructura factorial planteada, puede ser replicada, lo cual nos permitiría hablar de la validez factorial del instrumento.

Hay que aclarar al respecto que, en general, las estructuras factoriales pueden tender a ser inestables y dependen mucho del tamaño de la muestra. El tamaño necesario depende en parte de la cantidad de ítems involucrados en el análisis y hay diferentes criterios para establecer el número de sujetos indicado. Uno de los criterios, es que el número de sujetos no sea menor a cinco veces la cantidad de ítems del instrumento. Aún con este número de sujetos las estructuras factoriales pueden seguir siendo inestables, y no es sencillo replicar exactamente las mismas, máxime cuando el instrumento se aplica en un país o cultura diferente a aquella en que ha sido construido. Por tal motivo, consideramos, que al replicar la estructura factorial, deberíamos considerar la validez factorial, como una cuestión de grado, desde una replica perfecta, hasta la obtención de una estructura de los factores, por lo menos similar a la original. De todos modos, cuando la estructura factorial, no es exactamente igual, pero guarda suficiente similitud con la predicha, tenemos la opción de considerar que el instrumento tiene validez factorial y aplicarlo tal cual, o introducir modificaciones acordes con los datos obtenidos en el nuevo análisis factorial, tratando de preservar mayormente las características originales del test. Esta última opción nos parece la

Pag. 15-33 4/27/06 12:20 PM Page 30 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

más adecuada, principalmente si el test se está validando en un contexto diferente al de su construcción original y se dispone de los medios para hacer la adaptación correspondiente.

• **Diferenciación entre grupos:**
Se aplica el instrumento a dos o más grupos, los cuales debieran diferir en el atributo que se mide, en razón de alguna característica que se usó para formar los grupos, y que se podría inferir a partir de la teoría del constructo que mide el instrumento, que dicha característica estaría relacionada con diferencias predecibles en el atributo. Como ejemplifica Martínez Arias (1996): Si tenemos un test que mide inteligencia, teniendo en cuenta que las habilidades cognitivas varían con la edad, podríamos formar grupos de distintas edades, de acuerdo a las diferencias cognitivas que fueran esperables, y esperaríamos encontrar diferencias entre los grupos en cuanto al nivel de inteligencia que mide el test.

• **Correlaciones con otras medidas del constructo:**
Cuando ya existe otro test ya validado que mide el mismo constructo, o varios tests, se puede correlacionar el nuevo con aquél, o con los varios tests ya existentes, para establecer su validez de constructo. Este procedimiento, depende de que exista, por lo menos un test, con un constructo idéntico o muy similar, al que se intenta validar.

• **Las Matrices multimétodo-multirasgo:**
Para poder aplicar esta técnica se precisa que existan como mínimo dos métodos diferentes para medir el constructo que se va a validar. También se necesitan otros constructos que puedan ser medidos por los mismos métodos. Se miden en los sujetos de la muestra los distintos constructos con métodos diferentes. Se calculan las correlaciones entre todas las medidas y se forma con ellas una matriz que contendría los siguientes datos:

- Coeficientes de fiabilidad: serían las correlaciones obtenidas entre medidas del mismo constructo con el mismo método. Se espera que sean elevados.

- Coeficientes de validez convergente: son las correlaciones entre las medidas del mismo constructo, obtenidas con métodos diferentes. Se espera que las correlaciones sean altas.

- Coeficientes de validez divergente: son las correlaciones de constructos diferentes medidos con igual método y las correlaciones de diferentes constructos medidos con distintos métodos. Se espera que sean mucho más bajas que las obtenidas en la validez convergente y en el cálculo de la confiabilidad.

Pag. 15-33 4/27/06 12:20 PM Page 31 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

Conclusiones

La falta de confiabilidad de un instrumento, por lo general, afecta la validez de alguna forma, pero esto no significa que pueda tomarse entonces la confiabilidad como un indicador de validez, ya que un test podría tener muy alta confiabilidad y no por eso que esa medida sea válida, es decir, que esté midiendo el atributo que se espera. Tanto la validez como la confiabilidad, son importantes, porque es relevante que un instrumento sea lo más exacto posible, que contenga el menor número de errores; pero de qué nos sirve la precisión, que el instrumento tenga escaso margen de errores aleatorios, si lo que mide no es lo que suponemos, o sea, que no está midiendo la variable que realmente nos interesa medir. Por este motivo el test debe tener para ser adecuado, confiabilidad y también validez.

Consideramos que hay dos aspectos para pensar los análisis de validez: 1) desde un punto de vista teórico, o 2) con un sentido esencialmente práctico, focalizado en la utilidad del instrumento.

Desde un punto de vista teórico, nos interesa fundamentalmente la relación del instrumento con el constructo y las cuestiones relacionadas con la teoría que incluye al constructo, en cuanto a las hipótesis que pudieran derivarse de la misma para someter a prueba al instrumento mediante dichas hipótesis.

Si la validez la pensamos con un sentido práctico, no importaría tanto el constructo que se mide, como los resultados empíricos obtenidos y sus posibles aplicaciones. De esta manera podemos hablar de que un instrumento sea válido, pero también de que sea útil. Por éj, si un instrumento tuviera claro poder predictivo, que pudiera derivarse en aplicaciones útiles (establecidas a través de la validez de criterio); si no pudiera demostrarse con claridad su validez de constructo, ¿eso invalidaría su uso práctico? Seguramente que no.

En el caso del constructo, estaríamos en el orden de lo explicativo y esto sirve fundamentalmente para generar nuevas ideas y, para mejorar la capacidad de predecir, ya que la explicación, implica necesariamente la predicción. Ahora, como ocurre en la ciencia en general, podemos tener poder predictivo, en ausencia de poder explicativo y eso no quita que ese poder predictivo, nos permita generar aplicaciones útiles. Las investigaciones de validez nos proporcionan información, que al estar corroborada, nos permite avanzar hacia resultados con aplicaciones prácticas. Interesa establecer la validez relacionada con el constructo, pero sin perder de vista las investigaciones de validez que derivan en usos prácticos.

Podemos establecer un paralelismo, entre la investigación en psicometría y la inves-

Pag. 15-33 4/27/06 12:20 PM Page 32 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

tigación psicológica en general. Las formas y procedimientos de investigar en psicometría, no debiéramos considerarlos como absolutamente específicos de la misma. En general la psicometría encuadra en métodos de investigación generales, por éj., la confiabilidad por test- retest, puede ser equiparada a un diseño longitudinal de panel, la validez predictiva con diseños prospectivos y la concurrente con diseños transversales. En general la metodología usada es correlacional, pero no es la única posible. El constructo forma parte de teorías. De estas teorías se pueden derivar hipótesis, que al ser sometidas a prueba, aportan a la validez de criterio y/o a la de constructo, pero que en definitiva, también sirven para enriquecer el campo de conocimiento a que se refieren, más allá de la psicometría. En dichas investigaciones, al plantearse hipótesis funcionales, podrían haberse usado, tanto diseños experimentales, como diseños de grupos naturales, formados a partir de las medidas obtenidas en algún instrumento. O sea, que investigaciones realizadas con fines estrictamente psicométricos de análisis de validez, pueden aportar a diversas áreas de conocimiento, y al mismo tiempo, investigaciones no realizadas con fines psicométricos de validación, sino de investigación en áreas diversas de la psicología, pueden aportar a los estudios de validez, al elegir instrumentos psicométricos como formas de operacionalización de las variables en estudio. Por ej., si hiciéramos un estudio sobre como ciertos factores de personalidad pueden intervenir facilitando o dificultando la evolución terapéutica en determinada psicopatología, puede ser que la investigación se hiciera dentro de estudios sobre personalidad o en estudios clínicos y no como un análisis de validez, pero si en la operacionalización de la personalidad utilizamos una técnica psicométrica, el mismo estudio podría considerarse como un estudio de validez. No debemos olvidar que los tests operacionalizan variables y que estas formas de operacionalizar, son sumamente prácticas, motivo por el cual se utilizan en las áreas de investigación más diversas, al igual que en las más variadas aplicaciones, tanto clínicas, como en psicología laboral, en educación, etc. Esto da la razón de ser de la técnica, que en última instancia es creada para ser utilizada en investigación, o en distintas áreas de la psicología y crea un nexo indubitable entre la investigación psicométrica y la investigación psicológica en general.

Las distintas formas de validez, se encuentran muy relacionadas entre sí y, muchas veces, no diferenciadas claramente. Por ejemplo, una investigación sobre validez predictiva puede aportar a la validez de constructo, en la medida que la predicción pudo haberse deducido de la teoría que incluye al constructo. Consideramos entonces, que quizá, una diferenciación más clara, sea la que pase no tanto por las características de la investigación, sino por el propósito con que se la utilice. Si se la utiliza para analizar o justificar cuestiones teóricas relacionadas con el constructo, sería validez de constructo; si el propósito con el que se la realiza es establecer la utilidad del instrumento en aplicaciones prácticas o para corroborar hipótesis que vinculan el

Pag. 15-33 4/27/06 12:20 PM Page 33 ramon OMEGA:TRABAJOS 2006:b) febrero 2006:Libros Febrero 2006:Su

constructo con otras variables, sería validez de criterio.

Si bien en las técnicas psicométricas la validez y la confiabilidad son esenciales, no es menos cierto que, en cual de estas cuestiones o en que tipos de validez y confiabilidad se deba poner mayormente el énfasis, o en que elementos de los estudios se deba reparar especialmente y bajo que perspectiva, depende mucho, del instrumento que se vaya a utilizar, del propósito que se persiga y del contexto en que se piense emplear.

Por último, digamos que el uso adecuado de una técnica psicométrica, no se reduce a conocer como se aplica a los sujetos. Sino que fundamentalmente se trata de poder interpretar lo más correctamente posible las puntuaciones obtenidas. Para lo cual se requiere, conocer bien sobre validez y confiabilidad, en general y en particular sobre el test que se desee aplicar. Además de conocer lo más posible la teoría de la cual se derive el instrumento, lo cual nos va a permitir establecer relaciones teóricas del instrumento con otras variables relevantes y poder extraer de su aplicación la mejor información que redunde en variados y útiles usos prácticos.

Bibliografía

Arnau Gras, J. (1982). *Psicología experimental.* México: Editorial Trillas.

Eysenck, H. J. (1978). *Usted y la neurosis.* Buenos Aires: Editorial Abril.

Martinez Arias, R. (1996). *Psicometría: Teoría de los tests psicológicos y educativos.* Madrid: Editorial Síntesis.

Morales, P. (1988). *Medición de actitudes en Psicología y Educación. Construcción de escalas y problemas metodológicos.* San Sebastián: Editorial Ttarttalo en colaboración con la Universidad de Comillas.

Polit, D. F. y Hungler, B. P. (1997). *Investigación científica en Ciencias de la Salud* (5º ed). México: McGraw –Hill Interamericana.

Sampieri, R. H., Collado, C. F. y Lucio, P. B. (1997). *Metodología de la Investiga - ción.* Colombia: McGraw-Hill.

Fecha de recepción: 20 de agosto de 2005
Fecha de aceptación: 15 de octubre de 2005