11

Validez

1. Definiciones clásicas de validez

Se denomina validez de un test al grado en que el test realmente refleja aquello que dice medir. Diversos autores han expresado este concepto de un modo consistente, aunque con diferentes matices.

Gulliksen (1950, Cap. 9. pag. 88) había establecido que “La fiabilidad ha sido vista como la correlación de un test dado con una forma paralela. Correspondientemente, la validez de un test es la correlación del test con algún criterio. En este sentido el test tiene una gran cantidad de muy diferentes ‘valideces’”.

Para Lord y Novick (1968, Cap. 12, pag. 261) “Semánticamente, y en un lenguaje relativamente inexplícito, podemos definir la validez de un test como el grado en que mide o predice algún criterio de interés.”

Zeller y Carmines (1980, Cap. 4. pag. 78) definen validez como “que un instrumento mida exactamente lo que se supone que mide y nada más” Este ha sido un modo usual de presentar la cuestión.

De Gruijter y van der Kamp (1984, Cap.8, pag. 136) explican que “en un sentido muy general la validez de un instrumento de medida se refiere a lo que el instrumento está pretendiendo medir.”

Yela (1984, Cap. 10, pag 154) dice que “un test es válido si sirve para lo que con él se pretende”, y más abajo, explícitamente “son válidos [los tests] en la medida en que miden lo que dicen medir.”

Santisteban (1990, pag. 149) dice que “un test es válido si cumple su objetivo de realizar bien la medida de aquello para lo que se construyó.”.

Crocker y Algina (1986, Cap. 10, pag.217), -siguiendo una linea argumental en la que había abundado Cronbach (1971),- definen implícitamente validez como el grado en que las inferencias que se pueden hacer con las puntuaciones del test están justificadas.

2. Tipos de validez

Ha habido diversas clasificaciones de validez, más o menos solapadas y diversas especificaciones del concepto de validez que han dado lugar a varios tipos y subtipos específicos de validez.

Validez empírica y validez teórica

Lord y Novick (1968) distinguieron entre validez empírica y teórica. Definen la *validez empírica* como "el grado de asociación entre la medición y alguna otra medición observable". Se refieren a la *validez teórica* como una clase más amplia y significativa de validez que implica la relación de la medición observable "con algún constructo teórico de interés (variable latente)".

Según Lord y Novick la validez teórica incluye como un caso particular la *validez de constructo*, (a lo que añaden -sin más explicación- que "una validez teórica que no es necesariamente una validez de constructo es la raíz cuadrada de la fiabilidad del test").

Según Lord y Novick (1968, pags. 278-279) la *validez de constructo* es "el grado en que un test mide el constructo que fue diseñado para medir. [...] La dificultad para establecer la validez de constructo de un test es que el criterio, el constructo, no es medible directamente. De aquí que la correlación test-criterio no pueda calcularse." A continuación sugieren que, en su lugar, pueden calcularse las correlaciones del test con otros tests. El test debería correlacionar con aquellos que teóricamente deberían correlacionar (*validez convergente*). Y el test no debería correlacionar con aquellos tests con los que que teóricamente el constructo no debería correlacionar (*validez discriminante*). Si ambos conjuntos de hipótesis se cumplen puede afirmarse que el test presenta validez de constructo. Los términos validez convergente y discriminante habían sido introducidos por Campbell y Fiske (1959).

Validez criterial, validez de contenido y validez de constructo

La clasificación más ampliamente divulgada y aceptada distingue tres tipos de validez: la validez criterial (con dos variantes: la concurrente y la predictiva), la validez de contenido y la validez de constructo.

Esta clasificación es el fruto de un intento de consenso conceptual y terminológico desarrollado por una comisión de la APA *(American Psychological Association)* en los años 50, y ha sido elevada a convención por su presentación en sus *Standards for Educational and Psychological Tests* que son comúnmente bien aceptados como referencia para la medición en ciencias humanas.

Se denomina *validez criterial* a la que establece la relación entre un test y un criterio externo.

En general, se denomina *criterio* a cada variable relevante ajena al test con la que se espera que el test *se relacione* estadísticamente de determinada forma (generalmente, que *correlacione* de determinada forma).

Se denomina *coeficiente de validez* del test a la correlación entre test y criterio. Como se ha mencionado en la anterior cita de Gulliksen, esto significa que un test tiene tantos coeficientes de validez como criterios distintos con los que lo correlacionemos.

La validez criterial se denomina *validez concurrente* si el criterio se mide aproximadamente al mismo tiempo que el test.

La validez criterial se denomina *validez predictiva* si el criterio se mide tiempo después que el test (a veces años después).

(La validez criterial se denomina *validez retrospectiva* si establece la relación entre un test medido en el presente y un criterio que fue medido en el pasado. Este es un caso francamente poco frecuente, aunque teóricamente posible).

En términos de los *Standards* (1974), se denomina *validez de contenido* a la que muestra "cuan bien el contenido del test muestrea las clases de situaciones o cuestiones" que pretendidamente mide el test. Consiste en determinar si los items del test son una buena muestra que representa bien el *dominio de contenidos* que teóricamente pretende medir el test.

Los *Standards* (1974) establecen que "la *validez de constructo* se evalúa investigando que cualidades psicológicas mide un test." Según de Gruijter y van der Kamp (1984) la validación de constructo "es un proceso para desarrollar interpretaciones de observaciones (incluyendo registros de desempeño en tests) en relación a teoría psicológica o teoría de la conducta [...] implicando una variedad de investigaciones utilizando diferentes clases de métodos de análisis". La validez de constructo no tiene un método característico. Por el contrario, puede utilizar cualquier método que permita contrastar que las hipótesis que deberían seguirse si el instrumento mide el constructo que pretende medir, efectivamente se pueden afirmar.

Un "constructo" no es otra cosa que un concepto. Un concepto más o menos psicológico y más o menos mal elaborado teóricamente. Por ejemplo se denominan como constructos la inteligencia, cada factor de personalidad, las aptitudes, las actitudes, etc. Prácticamente la palabra constructo se utiliza para referirse a cada concepto teórico psicológico inobservable. Posiblemente el término es necesario para esconder el hecho de que, frecuentemente, no se sabe bien de qué se está hablando (si es que se está hablando de algo). Cuando se pregunta, por ejemplo, si un factor dado de personalidad es un proceso mental, tiende a

negarse. Si se pregunta si es un estado mental tiende a negarse. Si se pregunta si es un contenido mental, ó pensamiento, ó ... también se negará. Si se indaga si es conducta posiblemente tampoco pueda contestarse que lo es. Los constructos, como el alma, parecen no estar en parte alguna aunque parecen explicarlo todo. (A diferencia de ésta espero, por mi parte, que no tengan también vida eterna).

Originalmente la validez de constructo fue concebida especialmente para referirse a la validez de aquellos instrumentos de medida que pretenden medir inobservables psicológicos para los que no se encuentran criterios razonables ni hay forma tampoco de definir de un modo exacto el dominio de contenido ni, por tanto, el muestreo de contenido. Sin embargo, la generalidad de su formulación y la ausencia de métodos propios característicos, convierten de hecho, a la validez de constructo, en sinónimo de validez en general. La validez de criterio y la de contenido pueden verse como casos particulares, mejor acotados, de la validez de constructo. Esta generalidad del concepto puede verse por ejemplo en la definición de validez de constructo que hace Yela (1984, pag.160) “consiste en comprobar, según la metodología de la investigación científica, que el test mide efectivamente la variable a que se refiere”.

Existe un buen número de términos específicos más para designar algunos tipos concretos de validez, sin embargo no vamos a embarcarnos en su enumeración. Algunas de las denominaciones y aspectos principales pueden consultarse en Yela (1984) o Santisteban (1990) para citar textos españoles.

3. Relación entre fiabilidad y validez

Si en la fórmula de la correlación de Pearson que define la validez de un test X y un criterio Y se sustituyen X e Y por su descomposición en puntuaciones verdaderas y errores, en unos pasos se alcanza una fórmula conocida como *fórmula de atenuación*, expresión que puede considerarse una abreviatura para indicar que la fórmula expresa la correlación verdadera entre test y criterio una vez atenuados los errores de medida, o, si se prefiere, en otra expresión, cual sería el coeficiente de validez del test si tanto el test como el criterio hubiesen sido medidos por instrumentos perfectamente fiables. La fórmula es la siguiente:

$$r_{V_X V_Y} = \frac{r_{XY}}{\sqrt{r_{XX} \cdot r_{YY}}}$$

Puede seguirse la deducción de la fórmula y el estudio de una familia de casos semejantes en Yela (1984; Cap. 11, pág. 168 y siguientes) o en Santisteban (1990; Cap. 9; pág. 154 y siguientes).

La fórmula ha sido tenida por central en la teoría clásica desde que fuera deducida por Spearman en los primeros trabajos sobre el modelo de las puntuaciones verdaderas. Permite estimar la correlación entre las puntuaciones verdaderas de test y criterio; la correlación "verdadera" entre test y criterio, aquella que se daría, teóricamente, si no hubiera error de medida ni en el test ni en el criterio.

Permitir la estimación de la correlación real entre tests y criterios ha sido una de las principales justificaciones de la teoría de la fiabilidad para la Teoría Clásica de Tests.

En la realidad ni los tests ni los criterios pueden medirse de modo que su coeficiente de fiabilidad sea 1, es decir de un modo perfectamente fiable. El resultado de la fórmula es meramente teórico y especulativo. Por esta razón, desde un punto de vista aplicado tiene poco interés utilizar la fórmula de atenuación en lugar del coeficiente de validez correspondientes que expresa la correlación real entre las puntuaciones empíricas. *De hecho*, no disponemos de mediciones perfectamente fiables, -que no existen,- y por tanto, *de hecho*, nuestras puntuaciones, predicciones e inferencias empíricas e imperfectas son mejor descritas por la correlación corriente entre el test y el criterio.

Ejemplo de aplicación de la fórmula de atenuación

Supongamos un test con coeficiente de fiabilidad de 0'9 y un criterio con coeficiente de fiabilidad de 0'91. El coeficiente de validez de este test con este criterio es 0'7. ¿Cuál será la correlación entre test y criterio una vez atenuados los errores de medida?

Solución:

$$r_{V_x V_y} = \frac{r_{xy}}{\sqrt{r_{xx} \cdot r_{yy}}} = \frac{0'7}{\sqrt{0'9 \cdot 0'91}} = 0'773492$$

4. Concepción restrictiva de la validez

Existe un consenso justificado acerca de que la validez es el aspecto más importante de un instrumento de medición psicológica. Siendo así, la validez es el aspecto que más debería cuidarse en la elaboración de un instrumento. Y debería garantizarse que la validez es no sólo el mero cumplimiento formal de unos requisitos o la estimación mecánica de una fórmula. La concepciones o las concepciones tradicionales de la validez abarcan tanto procedimientos de prueba directa, tales como la relación con indicadores de rendimientos observables de los sujetos,

como procedimientos de prueba que se suponen indirectos, y cuyo información proviene del mismo instrumento o de otros que requieren las mismas necesidades de validación. De este modo, los procedimientos de validación tradicionales admiten cualquier clase de evidencia que pueda considerarse en el marco de la discusión de las hipótesis ligadas al constructo que el instrumento pretende medir. Esta generalidad tiene sin duda aspectos positivos, pero en mi opinión, falla en diferenciar la naturaleza de las fuentes de información que proveen la evidencia de validez, y, de ese modo, no diferencia adecuadamente aquellas fuentes de naturaleza objetiva, medidas de modo externo al test, relevantes y útiles en términos psicológicos y que son independientes del entramado de mediciones psicológicas expuestas a necesidades de validación. Esta clase de fuentes han de ser necesariamente de especial interés en conectar los instrumentos de medición psicológica y la vida real de los sujetos y merecen ser evaluadas de modo diferenciado identificando el grado de *validez restringida* que presenta el test.

Introduzco explícitamente el término de validez restringida para referirlo exclusivamente a la validez contrastada con esta clase de fuentes. Un test presenta validez restringida si permite pronosticar satisfactoriamente variables *observables* y *relevantes ajenas* al test. Y es valido, concretamente, *para* pronosticar *esas* variables y en las condiciones determinadas en que *se haya probado* que es capaz de hacerlo.

La validez restringida es pues la capacidad de pronosticar variables observables y relevantes concretas ajenas al test: La capacidad asociada a un test de producir inferencias correctas sobre dimensiones observables de la realidad.

En mi opinión la validez restringida es el núcleo básico de la validez psicológica de un test, una clase de validez irrenunciable y sin la cual cualquier otra puede ser irrelevante en términos de utilidad psicológica práctica.

5. Algunas consecuencias de la definición de validez restrictiva

La definición de validez anterior es intencionadamente restrictiva y abiertamente polémica respecto a algunas de las definiciones al uso y, desde luego, respecto a muchos de los usos actuales.

La definición de validez restrictiva admite únicamente criterios observables y relevantes, ajenos al test. Esta afirmación comporta limitaciones importantes acerca de qué variables pueden admitirse como criterios.

La definición de validez restrictiva excluye explícitamente como criterios adecuados a los demás tests, indicadores de otras variables inobservables. Excluye como criterios

incluso a los demás tests que se consideran validados por pruebas anteriores.

Generalmente se admite que la validez de un test consiste en el *grado en que mide aquello que debe medir y no mide ninguna otra cosa*. Esta concepción es de significado bastante transparente para algunas dimensiones físicas donde precisamente no parece que exista esta clase de problema. Parece, por ejemplo, que un metro solo puede medir longitud (aunque, de hecho, podría medir otras cosas como temperatura, vía efecto de dilatación). Pero la cuestión es más complicada en algunas partes de la psicología donde se pretende medir supuestos constructos psicológicos inobservables (muchas veces de dudosa entidad).

¿Cómo saber que un test está midiendo el constructo inobservable A? El modo tradicional implica razonamientos y procesos como los siguientes. Primero, se define el constructo A. Segundo, se establecen teóricamente las respuestas observables asociadas a A. Tercero, se formulan items o cuestiones que se supone suscitan respuestas asociadas a A, y se dice que esas respuestas asociadas a A son una medida de A. Cuarto, se administran a una muestra en la que se espera que A este presente en diferentes grados. Quinto, si alguna de las presuntas respuestas asociadas a A no correlaciona con las otras es excluida de la medición de A en el proceso de elaboración del instrumento de medida de A. Sexto, se explora el modo en las respuestas se agrupan y si es necesario se remodela el concepto de A y su estructura para ajustar razonablemente con el comportamiento empírico de las supuestas respuestas asociadas a A. En todo caso esta remodelación se considera basada en datos experimentales de modo que ¿cómo podría estar equivocada? Pueden añadirse algunas sofisticaciones más pero esencialmente, el ciclo está completo: Acaban de nacer el constructo A y su test A, simultáneamente. ¿Cuál es la entidad del constructo A? ¿Mide el instrumento de medida de A realmente a A? Con este tipo de procedimientos, inevitablemente, incluso si A carece de entidad.

En ocasiones para validar A se hipotetiza que el nuevo instrumento sobre el inobservable A correlaciona con el del inobservable establecido B. Al desarrollar el instrumento de medida de A tenemos en cuenta que los items correlacionen con el total de A (que todo sea A), y que correlacionen también con B, como hipotetizábamos. Al acabar de construir A lo correlacionamos con B. Entonces descubrimos que A correlaciona más o menos con B. Perfecto, hipótesis satisfecha. Ya tenemos dos: A y B. Las correlaciones no son nunca perfectas (por lo general están sustancialmente lejos de +1 ó -1. Y, además, no hay criterio claro acerca de cuan altas deben ser. No obstante por lo general se admiten las correlaciones que se hayan obtenido como expresión de la relación entre A y B.

La significación estadística de la correlación entre tests y criterio tal como se utiliza convencionalmente no tiene entidad para justificar determinada validez: Con un tamaño de muestra adecuado, dos tests pueden correlacionar de modo estadísticamente significativo aunque la correlación sea teóricamente insignificante y poco útil para el pronóstico. Por otra parte el uso convencional es muy laxo respecto a la magnitud de las correlaciones esperadas y, ante la insuficiencia teórica los análisis se limitan a contrastar si la correlación difiere significativamente de 0, una clase de hipótesis de poca sustancia teórica en muchos casos.

Si se repite el proceso una y otra vez, con un poco de esfuerzo, se alcanzan los literalmente centenares de constructos psicológicos definidos (o pseudo-definidos) y sus centenares de instrumentos de medida. Muchos de ellos "validan" a otros muchos en cadenas de relaciones donde unos psicólogos le dan la razón a otros (serán amigos).

Convencionalmente, un instrumento nuevo (y su "constructo") se justifican bien si correlacionan con otro (o mejor otros) ya aceptados. Por cierto que si son función lineal de aquellos ¿para qué sirven? y si no lo son ¿cómo se usa un coeficiente de correlación para validarlos? Si no se definen las hipótesis de validez con precisión todo valor entre ambos extremos puede parecer aceptable.

En mi opinión la teoría y los usos convencionales son demasiado poco estrictos con la naturaleza de los criterios a utilizar. Creo que los criterios substanciales de un test han de referirse a resultados tangibles, observables, relevantes y ajenos a la maraña de conceptos psicológicos entrelazados misma denominada a veces "*red nomológica de constructos*", aunque esta denominación sofisticada no añade un gramo de certeza.

Esta orientación a lo observable no ha sido el punto de vista favorito en el desarrollo de las mediciones psicológicas. Puede argumentarse en su contra que podemos estar legítimamente interesados en "medir" experiencias personales subjetivas e intransferibles (por ejemplo, sensación de bienestar psicológico o sensación de tensión) Por supuesto parece que la psicología debe ocuparse de este difícil terreno, genuinamente propio. Pero analicemos un punto más la cuestión. Estas variables o "constructos" (como es costumbre llamarlas) pueden ser de dos tipos. O bien son *distintas e independientes* de criterios externos observables (**variables tipo a**). O bien *están asociadas* como causas, como efectos o como covariadas a criterios externos observables (por ejemplo, conductas, rendimientos, resultados prácticos, etc.; **variables tipo b**).

Para las *variables tipo a,* todos los procedimientos que podamos sugerir para evaluar si un instrumento mide fielmente un constructo de este tipo han de estar basados en la subjetividad. Finalmente, el constructo será lo que

diga un psicólogo que sea, porque no hay forma de decir que no es así. A lo sumo podremos sentar juntos a diversos psicólogos y pedirles que nos digan si les parece bien o no lo que el primero ha dicho. Esto es lo que se denomina criterio de "jueces" o "expertos". Hay métodos sofisticados para realizar esta clase de consultas y evaluar el acuerdo entre los jueces. Generalmente el acuerdo entre los jueces será menos que perfecto y acabaremos decidiendo cual de las subjetividades nos parece mejor. Hay modos excelentes de envolver esto en papel de regalo: por ejemplo calculando porcentajes de acuerdo o correlaciones entre juicios y otros estadísticos, pero seguimos dentro de la subjetividad. *Ningún método estadístico o psicométrico aplicado sobre información subjetiva produce por sí información objetiva.* La objetividad es una cualidad de la fuente y del proceso de obtención de información, que no puede ser suplida en el proceso de análisis. En el mejor de los casos un método de expertos, y los hay cuidadosos y sofisticados, solo permite concluir que los expertos están de acuerdo, no que una proposición sea cierta o no. La verdad científica no es una cuestión de popularidad democrática, ni siquiera entre expertos.

Con las *variables tipo a* se produce un efecto impresionante: Diga lo que diga el experto, si no se dispone de un referente objetivo ¿cómo podemos afirmar que la experiencia subjetiva que nos trasmite una persona, por ejemplo, de sensación de sufrimiento psicológico, es o no es cierta y en que grado lo es? Si hablamos de una *variable de tipo a* para la que no aceptamos indicadores objetivos creo que esta clase de problemas siguen siendo esencialmente irresolubles.

Desde la definición de validez restrictiva que he planteado para estas variables podríamos hablar de *pseudo-validación* (por ejemplo por métodos de jueces) pero no de una validación auténtica.

En el caso de las *variables tipo b* se puede estudiar la variable y su instrumento de medida a través de sus relaciones con indicadores objetivos. Puede hablarse de validación según la definición restrictiva inicial. La cuestión esencial es que si una variable es de tipo b, es decir, presenta relaciones con indicadores objetivos observables, externos y ajenos al test, entonces la validación de un instrumento de medida de la misma *debería* basarse de modo sustancial en conocer perfectamente esas relaciones. Los procedimientos pseudo-validatorios, como opiniones de expertos, métodos factoriales o relaciones con otras variables que no son indicadores objetivos observables externos y ajenos al test. Estos procedimientos pseudo-validatorios tienen en mi opinión un papel secundario. Pueden tener un interés práctico, y puede tener sentido utilizarlos para comprender el comportamiento de la variable y su relación con otras de interés, pero no sustentan una validación en el sentido fuerte de la validez restrictiva.

Alguien podría argumentar que puede existir un caso c, en el que constructos inobservables estén asociados a constructos inobservables. Este falso nuevo caso contribuye esencialmente a la actual *Babel de los constructos psicológicos,* así que conviene despacharlo claramente. Veamos. Un constructo inobservable está asociado a otro u otros inobservables. Bien, entonces, o bien estos otros no están conectados a criterios empíricos (caso a) o bien lo están (caso b). Si no lo están entonces el constructo inicial tampoco lo estará y el problema se reduce a un caso a. Si los otros constructos sí están asociados a indicadores empíricos entonces el constructo de interés también lo estará y, en principio, deberíamos poder tratarlo como una variable tipo b. Puede argumentarse aun que podría estar asociado a otros constructos asociados a variables observables sin estar asociados a variables observables. Bien, entonces podemos seguir tratándolo como un caso a, pero, mejor todavía ¿Por qué no nos interesamos directamente por esos otros constructos?

La cuestión de constructos inobservables que llaman a constructos inobservables y así hasta el infinito va en contra del elemental principio de la parsimonia científica. No hay motivo para sostener entidades inobservables cuya existencia no esté exigida por datos de la experiencia. No se debe multiplicar los constructos sin necesidad. Pero no deberíamos tener que mencionar estas cosas, ésta era una discusión de actualidad en torno al 1330, en tiempos del franciscano Guillermo de Ockham. Sin embargo, en psicología, parece que periódicamente es necesaria una relectura de aquellas cuestiones. Por ejemplo, los nuevos modelos cognitivos que tratan de explicar las respuestas de los sujetos ante los tests psicométricos pueden también estar expuestos en muchos casos a esta clase de críticas.

La definición de validez restrictiva excluye los procedimientos subjetivos de validación, tales como aquellos que se basan en jueces. También excluye los procedimientos basados en la subjetividad, los basados en relaciones internas de la especie items-factor o items-rasgo latente, y, en general, todos aquellos que no se basan en datos que se puedan considerar objetivos, observados directamente, relevantes operativamente y tomados de modo independiente y ajeno al test. Desde el punto de vista de la validez restrictiva estos procedimientos podemos considerarlos globalmente en la categoría de métodos pseudo-validatorios.

Obsérvese que buena parte del arsenal de métodos para evaluar la validez que se han elaborado pueden ser pseudo-validatorios o de validez restrictiva dependiendo del uso que se haga de ellos. Por ejemplo, el simple y fundamental coeficiente de validez pertenecerá a una u otra categoría en función de la elección del criterio al que se refiera.

En mi opinión podemos escalar variables psicológicas subjetivas por procedimientos de jueces, pero me parece un círculo vicioso validar por procedimientos de jueces. Generalmente cuando un test se refiere a un constructo inobservable los items son elegidos por (al menos) un juez, el constructor de la escala, que los crea, analiza y ordena según las hipótesis que sustenta sobre el constructo y su instrumento de medida. Validar implica probar que estas hipótesis son validas. En mi opinión, la opinión del mismo juez o de otros, más o menos expertos, no es una prueba científica, excepto de su opinión.

La definición restrictiva también excluye como criterio cualquier variable cuya medición esté en algún modo relacionada con la efectuada por el mismo test. Por ejemplo, estimaciones subjetivas de frecuencias realizadas por los sujetos acerca de variables relacionadas con el test. De acuerdo con la definición restrictiva que he propuesto no puede, por ejemplo, diseñarse un conjunto de tres medidas subjetivas de un mismo constructo inobservable (por ejemplo, una medida del tipo escala de verdadero-falso, otra del tipo escala de elección entre alternativas y otra del tipo completar frases) y pretender que se validan entre sí. Podremos establecer que son consistentes entre sí pero no que se validan. Un conocido tenía dos termómetros corporales tradicionales. Cada vez que se encontraba mal utilizaba uno, y, para más seguridad, después el otro. Ambos estaban de acuerdo y, frecuentemente, daban más de 38 grados, por lo que, bien confirmada y afianzada su sensación subjetiva, realizaba frecuentemente tratamientos antibióticos, de acuerdo con su médico. Así fueron las cosas hasta que la industria japonesa llenó las farmacias de termómetros electrónicos. Al contrastar aquellos termómetros con estos otros se pudo comprobar que ambos termómetros tradicionales medían consistentemente entre si, pero sobreestimando la fiebre aproximadamente un grado. Muchas febrículas habían sido tratadas como fiebre y posiblemente algunos o muchos tratamientos antibióticos no estuvieron nunca justificados. En psicología la consistencia entre dos medidas sin referencia externa no puede admitirse como una prueba de validez, ni mucho menos como una prueba de la entidad del constructo.

Explícitamente la validez restrictiva excluye el concepto de "validez factorial" en función del cual un test es valido en la medida en que satura en (correlaciona con) un factor determinado congruente con la teoría. Esta es una discusión compleja que requiere atención a parte. Implica el tema del papel del análisis factorial confirmatorio y exploratorio en el contraste de hipótesis acerca de constructos inobservables. Los métodos factoriales son técnicas estadísticas útiles para comprender como comparten varianza conjuntos más o menos grandes de variables en muestras todavía más grandes. Buena parte

de la confusión teórica existente en psicología proviene de la falacia metodológica de identificar esos fragmentos de varianza compartida, generalmente resumidos en unas funciones lineales denominadas factores, como rasgos latentes con entidad psicológica. Esta clase de métodos provee factores allí donde haya cierta variabilidad compartida, independientemente de la naturaleza de esa covariabilidad. Esto significa que si existe una cierta asociación entre un conjunto de respuestas el método es capaz de articular esa varianza compartida en la forma de una nueva variable matemática función de las anteriores. Pero eso no indica en forma alguna entidad psicológica. El método factorial y su resultado no constituyen en modo alguno evidencia empírica que exija la existencia del constructo. Esta confusión entre la versatilidad y habilidad de un método para proveer funciones que integran porciones de varianza y la evidencia empírica que puede exigir un nuevo constructo ha contribuido esencialmente a multiplicar los constructos sin necesidad.

La validez restrictiva explícitamente excluye el concepto de "validez aparente" relativo al grado en que el test "parece" que mide lo que mide a los sujetos que han de utilizarlo como examinados, como interpretes de sus puntuaciones o como jueces que toman decisiones sobre las personas acerca de esas decisiones. En determinadas circunstancias que el test sea aparentemente valido para los examinados o para otros (por ejemplo directores de colegios, padres, abogados o magistrados implicados en decisiones sobre personas examinadas con tests) es fundamental para su buen uso. Así es. Pero, primero, esto no me parece razón para volver a introducir la opinión -en este caso ni siquiera de expertos- en el concepto de validez. Y, segundo, probablemente los constructores y usuarios de tests jamás hubieran tenido que enfrentarse a procesos judiciales y problemas con la validez aparente de sus tests si los tests fueran realmente válidos en el sentido fuerte y contrastado que se postula. La importancia de la validez aparente es parcialmente fruto de una carencia de validez restringida bien contrastada.

La validez restrictiva no se limita a la validez criterial tradicional con su énfasis en el coeficiente de validez pero aplicada sobre criterios objetivos externos y bien definidos.

En la validez restrictiva se mantiene abierto el estudio de las relaciones entre tests y criterios a cualquier tipo de relación funcional o causal, independientemente de que pueda expresarse o cuantificarse adecuadamente con un coeficiente de correlación. No está justificado técnicamente que la teoría de la validez criterial este tan pegada al coeficiente de correlación de Pearson como estadístico para definir la relación práctica esencial test-criterio. Primero, no hay porque sostener por sistema que tests y criterios se relacionan linealmente, por tanto es posible que modelos no lineales sean más explicativos. Esta posibilidad,

aunque rara vez contrastada, es una sospecha plausible en muchos campos cuando se comienzan a analizar datos reales.

Para un modelo de relación test-criterio bivariado en muchos casos será más útil, preciso y explicativo en términos psicológicos, confeccionar y utilizar una tabla de contingencia donde se aprecien las distribuciones del criterio Y condicionadas a los valores del test predictor X. Esto permite hacer pronósticos sustancialmente prácticos y con bien pocos supuestos.

Por ejemplo, conocer una tabla de contingencia entre un test de aptitud y un criterio de tiempo en el puesto de trabajo, permite predecir a partir de una puntuación X en el test de aptitud que existe una probabilidad p de ascender antes de a años.

Si hay que trazar la regresión del criterio sobre el test será una buena idea al menos explorar que forma describe esa regresión antes de forzarla a adoptar la expresión de una función afín.

Todavía si la relación va a ser descrita por la regresión lineal minimocuadrática habrá que analizar un conjunto de posibilidades y dificultades: valores extremos, papel de los datos faltantes, variables moduladoras continuas y discretas, etc.

Es posible que estemos interesados en manejar modelos más complejos con múltiples criterios y para los cuales habrá que utilizar las técnicas estadísticas oportunas. Un modelo relativamente sencillo es la regresión lineal múltiple, pero existen otros muchos que pueden ser de interés. Los modelos Lisrel (también llamados modelos de ecuaciones estructurales o modelos de estructuras de covarianza) se vienen usando con ventaja en aplicaciones más o menos sofisticadas en los últimos años. Son una metodología prometedora que también exige ciertas cautelas. En cualquier caso no hay razón para preferir el coeficiente de correlación a cualquier otro estadístico que resulte adecuado para describir la relación entre un test y un o unos criterios.

6. Consideraciones generales sobre la importancia de la validez

Creo, como muchos otros, que la validez es el aspecto más importante que podemos evaluar de un instrumento de medición psicológica. Pero, como hemos visto, no existe la "validez" de un instrumento en abstracto. *Cada instrumento ha de ser validado para cada uso concreto.* Es decir, cada interpretación, predicción o decisión basada en las puntuaciones de un test debe haber sido probada antes. No pueden sostenerse decisiones porque "parece" que se desprenden del test, o de su significado.

No hace mucho tiempo un semanario "rosa" sacaba en sus páginas cada semana un pequeño "test" sobre diversas cosas. Por ejemplo "¿Es usted un buen padre?, o "¿sabe Usted conquistar a las personas del otro sexo?". Cuestiones todas ellas de la máxima relevancia y utilidad práctica (aparentemente bastante más que el factor M del 16PF llamado "praxernia-autia", o el I, llamado "harria-premsia" sin ir mas lejos). Las preguntas eran razonables, quiero decir aparentemente no más estúpidas que las de cualquier test de personalidad o de hábitos sociales. Después, además, el semanario incluía una tabla para interpretar las puntuaciones: "De 0 a 10 puntos: Lo tiene fatal...[...] De 90 a 100 puntos: Fantástico, es Usted fabuloso..." Las tablas de interpretación también eran de bastante sentido común, o al menos a mi me lo parecían. Y desde luego mucho más informativas que algunos de los baremos inadecuados de los que hablaba páginas atrás.

Entonces, en medio de tanta serenidad intelectual ¿qué tenían de malo estos "tests" para no guardarlos presto en el archivo de tests psicológicos? (Así como uno guarda las fichas sobre como quitar las manchas de huevo o de sangre, por lo que pueda pasar). La cuestión es, simplemente, que *no se mencionaban pruebas empíricas que sostuvieran las interpretaciones* que se proponían. A mi me parecía, como opinión de experto en psicología, que, efectivamente, sacando allí menos de 10 puntos el examinado no conquistaría a nadie ni por asomo. Y no es una cuestión de falta de acuerdo entre jueces. Sin ir mas lejos unas amigas mías expertas en estas lides eran de mi misma opinión. El problema es que no había ni una prueba empírica mencionada de que, efectivamente, los sujetos que hubieran sacado en la prueba menos de 10 puntos disfrutarían de una vida célibe sin la incomodidad que acarrean las conquistas -y los riesgos sociales asociados.- Este pequeño detalle es la validez. Si hubieran pruebas que avalaran las interpretaciones de las puntuaciones estaríamos delante de un test, en su ausencia solo tenemos un divertimento.

Un test vale exactamente lo que valen sus estudios de validación. Con un test se pueden hacer exclusivamente las afirmaciones que sus estudios de validación demuestran que se pueden hacer.(Estas dos últimas afirmaciones me parecen de las más importantes que hay en todo este texto).

Si no se sabe si un test puede o no permitir hacer determinada inferencia hay que diseñar un estudio de

validación para ponerlo a prueba (antes de empezar a hacer la afirmación gratuitamente). La especulación es inevitable en ciencia y es inevitable en la práctica profesional, nos sirve para generar hipótesis y para razonar posibles cursos de acción. Pero debe distinguirse finamente lo especulativo de lo contrastado, y un instrumento de medida puede admitirse como tal en el grado en que se disponga de interpretaciones contrastadas para sus puntuaciones.

En psicología son frecuentes las relaciones muy de sentido común que el público y muchos psicólogos se creen a pies juntillas porque suenan bien, pero que *se ha demostrado* que son falsas, que no pueden sostenerse. Esas relaciones, prejuicios y juicios de sentido común que parecen ciertos pero que no lo son se denominan *correlaciones ilusorias.*

No se puede coger un test y empezar a hacer afirmaciones más allá de las que estrictamente se ha probado que se pueden hacer. Por eso un test no es nada sin el manual del test que es donde, junto a la normas de aplicación e interpretación, deben estar bien detallados y especificados todos los estudios que permiten sostener cada una de las afirmaciones que se pueden hacer con sus puntuaciones.

Una interpretación muy razonable que aun no ha sido puesta a prueba es una hipótesis. Solo una hipótesis. Hay que contrastarla antes de usarla. Los profesionales que utilizan tests como parte de su trabajo tienen frecuentemente la ocasión de realizar, con poco esfuerzo adicional más, estudios de validación locales de las relaciones más importantes para ellos.

> Por ejemplo, la primera vez que se seleccionan personas para una cierto puesto no hay más remedio que seleccionar los predictores (generalmente algún o algunos tests entre ellos) con criterios racionales amparados en los resultados obtenidos por otros colegas y publicados en las revistas científicas y en los manuales. Pero una vez que esas personas están trabajando es posible poner en relación tests y criterios relativos al trabajo (véase por ejemplo Santisteban (1990 pág. 175) o Yela (1984, Cap. 13). La próxima vez que se vaya a seleccionar personal podrán haberse adoptado correcciones en el procedimiento selectivo. Situaciones relativamente semejantes suceden con los tests de orientación profesional, con los que orientan hacia determinados tratamientos o con los que describen al sujeto en términos de patologías. Todos pueden ser contrastados con variables externas en el trabajo real si se hace el esfuerzo de plantearse su estudio desde el principio. Esto permitirá incluso verificar si algunas de las inferencias que según el manual se podían hacer realmente se cumplen en nuestros casos. La curiosidad profesional por

> descubrir el alcance de la eficacia de los propios diagnósticos y de las propias intervenciones debe formar parte del carácter de la psicóloga o del psicólogo aplicado. Esa curiosidad esencial reclama contrastar la validez de nuestros instrumentos, en un sentido práctico, inmediato, tanto como sea posible.

Este enfoque de la validez supone que ésta se pueda definir como el conjunto de hipótesis que han sido contrastadas (y que han recibido apoyo) para un test dado. Se puede definir la validación pues como un *proceso de contraste de hipótesis* en el que se establecen las interpretaciones aceptables de las puntuaciones de un test.

Si establecer la validez de un test es contrastar como hipótesis cada una de las interpretaciones, pronósticos e inferencias que deseemos hacer con él, entonces procede tener en cuenta todo el aparato estadístico y de diseño relativo al contraste de hipótesis.

Habrá que definir adecuadamente las hipótesis teóricas y traducirlas en hipótesis estadísticas. Habrá que analizar qué diseño, nivel de confianza, potencia y tamaño de la muestra deben utilizarse para qué diseño de análisis y qué prueba estadística, considerando la naturaleza de las variables y los objetivos de la investigación. Habrá que considerar y minimizar los factores que puedan afectar a la validez interna y externa del diseño. Habrá, en fin, que considerar todos los aspectos teóricos y prácticos que hay que considerar cuando se define y lleva a cabo una investigación.

Quizás el lector estará pensando que sería mejor adquirir una prueba comercial que dé todo esto hecho. Sin embargo, me temo que aun en este caso, contando con una buena prueba publicada, respaldada por buenos estudios de validación, puede ser necesario en algunos casos tener que probar algunas hipótesis adicionales para fines particulares o para una población particular.

En cualquier caso el proceso de validación de un instrumento de medición psicológica no termina nunca.

Primero porque, por lo general, no es suficiente con disponer de un solo estudio por lo que respecta a las características, inferencias o pronósticos más importantes que pueden efectuarse con el test. Los buenos tests tienen varios, y a veces muchos, estudios para respaldar su funcionamiento y su interpretación. Lo usual es validar y, al menos, volver a validar. Este proceso de validar de nuevo en una segunda muestra se designa a veces como *validación cruzada*.

En segundo lugar, los estudios de validación, como los baremos, habrían de presentarse con “fecha de consumo preferente” porque, efectivamente, tienen una caducidad temporal después de periodos más o menos largos. Esto se debe principalmente a que los tests suelen presentar un alto grado (o al menos un cierto grado) de contenidos

culturalmente dependientes. Por tanto los estudios necesitan ser revisados porque los sujetos, la cultura y los criterios pueden haber cambiado de diversas formas que afecten al test.

Lo ideal, cuando ello es posible, es ir acumulando evidencia a medida que se va trabajando con un test, de forma que todas o algunas de las inferencias de nuestro interés puedan ser reevaluadas periódicamente con nuestros propios datos.

En síntesis: *Nada más importante que la validez*; pero no la validez en general, sino la *validez específica para las interpretaciones, pronósticos o inferencias* que nosotros hacemos en concreto a partir de las puntuaciones del test.