TEMA 6. LA VALIDEZ

1. Concepto de validez

2. Validez de contenido

3. Validez de criterio o criterial
    - Validez externa / validez interna
    - Validez concurrente / validez predictiva

4. Validez de constructo
    - Métodos más utilizados

5. Cuestiones relativas a la estimación de la validez

6. Algunas amenazas a la validez de la investigación

Bibliografía

CONCEPTO DE VALIDEZ

La validez de un test indica el grado de exactitud con el que mide el constructo teórico que pretende medir y si se puede utilizar con el fin previsto. Es decir, un test es válido si "mide lo que dice medir". Es la cualidad más importante de un instrumento de medida. Un instrumento puede ser fiable pero no válido; pero si es válido ha de ser también fiable.

Se puede decir, que la validez tiene **tres grandes componentes**:

- *Validez de contenido*
- *Validez de criterio o criterial*
- *Validez de constructo*

Las tres se refieren a aspectos diferentes y la utilización de uno u otro concepto de validez depende del tipo de test.

1. VALIDEZ DE CONTENIDO

Se refiere al grado en que el test presenta una muestra adecuada de los contenidos a los que se refiere, sin omisiones y sin desequilibrios de contenido.

La validez de contenido se utiliza principalmente con tests de rendimiento, y especialmente con los tests educativos y tests referidos al criterio. En este tipo de tests se trata de comprobar los conocimientos respecto a una materia o un curso.

La validez de contenido descansa generalmente en el *juicio de expertos* (métodos de juicio). Se define como el grado en que los ítems que componen el test representan el contenido que el test trata de evaluar. Por tanto, la validez de contenido se basa en (a) la definición precisa del dominio y (b) en el juicio sobre el grado de suficiencia con que ese dominio se evalúa.

A pesar de que no se utiliza un índice de correlación para expresar la validez de contenido, existen procedimientos para cuantificarlo. Algunos de estos procedimientos son:

a) **Cálculo de descriptivos:** Consiste en calcular la media y la desviación típica de todos los ítems. A continuación, se determinará una puntuación de corte (o índice de validez de contenido) que refleje, en base a la evaluación de los expertos (método de juicio), que la puntuación del ítem es demasiado baja en relevancia como para ser incluido en la escala. No hay reglas. La decisión dependerá del juicio propio. Por este motivo, se han de tener en cuenta las siguientes cuestiones:

   - Ser capaz de justificar la decisión sobre la puntuación corte.
   - Usar la misma puntuación de corte para todos los ítems de la escala, incluso si la escala es multidimensional.
   - No poner una puntuación de corte tan alta que implique eliminar muchos ítems. (Se podrán eliminar más ítems con el cálculo de la fiabilidad y el análisis factorial).

b) **Índice de validez de contenido (IVC):** Lawshe (1975) propuso un índice de validez basado en la valoración de un grupo de expertos de cada uno de los ítems del test como innecesario, útil y esencial. El índice se calcula a través de la siguiente fórmula:

$$IVC = \frac{n_e - N/2}{N/2}$$

Donde n_e es el número de expertos que han valorado el ítem como esencial y N es el número total de expertos que han evaluado el ítem.

El IVC oscila entre +1 y -1, siendo las puntuaciones positivas las que indican una mejor validez de contenido. Un índice IVC = 0 indica que la mitad de los expertos han evaluado el ítems como esencial. Los ítems con una bajo IVC serán eliminados. Lawshe (1975) sugiere que un IVC = .29 será adecuado cuando se hayan utilizado

40 expertos, un IVC = .51 será suficiente con 14 expertos, pero un IVC de, al menos, .99 será necesario cuando el número de expertos sea 7 o inferior.

2. VALIDEZ DE CRITERIO O CRITERIAL

Se refiere al grado en que el test correlaciona con variables ajenas al test (criterios) con lo que se espera por hipótesis que debe correlacionar de determinado modo. Un ***criterio*** es una variable distinta del test que se toma como referencia, que se sabe que es un indicador de aquello que el test pretende medir o que se sabe que debe presentar una relación determinada con lo que el test pretende medir. Se denomina ***coeficiente de validez*** a la correlación del test con un criterio externo.

La elección del criterio es el aspecto crítico en este procedimiento de determinación de la validez, ya que es muy difícil obtener buenos criterios. Un mismo test puede tener más de un tipo de validez, es decir puede estar validado con respecto a varios criterios y los diferentes coeficientes de validez que resultan pueden tener valores diferentes.

Dentro del concepto de validez de criterio cabe distinguir a su vez entre:

- *Validez externa y validez interna*
- *Validez concurrente y validez predictiva*

a) Hablamos de ***validez externa*** si el test se ha validado con respecto a un criterio externo, como por ejemplo, una evaluación de rendimiento. La correlación del test con el criterio da lugar *al coeficiente de validez externa*. Hace referencia a la posibilidad de generalización.

   Sin embargo, hablaremos de ***validez interna*** si se correlaciona un test con otro con validez reconocida que mide el mismo rasgo; los *coeficientes de validez interna* suelen ser menores que los de validez externa y su interpretación es difícil. Para evitar errores de interpretación se suele correlacionar un test con todos los tests ya validados que miden lo mismo y calcular un *coeficiente de correlación múltiple*. Este coeficiente de validez interna suele alcanzar el valor del coeficiente de validez

externa. Hace referencia a la validez del resultado de la investigación para los sujetos estudiados.

b) La distinción entre validez concurrente y predictiva se emplea según se utilice un criterio disponible en el momento (***validez concurrente***) o cuando se pretenda predecir la conducta futura de un individuo (***validez predictiva***). Este tipo de validez se exige especialmente para los instrumentos que se utilizan en selección y orientación académica o profesional.

Cálculo del coeficiente de validez

Los procedimientos estadísticos utilizados en la validación referida a un criterio varían según el número de predictores utilizados (uno o más tests) y el número de criterios empleados (criterio único y criterio compuesto o múltiple). Martínez Arias (1995) distingue los siguientes casos:

1. *Un único test y un solo criterio*: se emplearían los procedimientos de correlación y regresión lineal simple.
2. *Varios predictores (tests) y un solo criterio*: se emplea la correlación y regresión lineal múltiple o el análisis discriminante.
3. *Varios predictores y varios criterios*: regresión lineal multivariante y la correlación canónica.

4. VALIDEZ DE CONSTRUCTO

Es un concepto más complejo. Se refiere al grado en que el instrumento de medida cumple con las hipótesis que cabría esperar para un instrumento de medida diseñado para medir precisamente aquello que deseaba medir. Se puede considerar un concepto general que abarcaría los otros tipos de validez.

El término ***constructo*** hace referencia a un concepto teórico psicológico inobservable (ej. la inteligencia, cada factor de personalidad, las aptitudes, las actitudes, etc.) La definición operativa de estos constructos presenta considerables dificultades en

la práctica, ya que no son directamente observables. Debido a esto, la validación de un constructo es un proceso laborioso y difícil.

Para la estimación de la validez de constructo se utiliza una metodología variada. Algunos de los **métodos más utilizados** son:

1) Métodos correlacionales

Los coeficientes de correlación nos indican la relación del test con el conjunto de instrumentos de medida y criterios posibles, así como la relación entre el test y el constructo.

- Correlación del test con un criterio externo
- Correlación test con otros tests que pretenden medir los mismos aspectos o aspectos semejantes.
- Correlación del test con otros tests que miden características, que nada tienen que ver con el constructo que subyace al test.

Campbell y Fiske (1959) proponen que se calcule dos tipos de validez:

a) *Validez convergente*: indica las correlaciones positivas con otros tests que miden lo mismo.
b) *Validez discriminante*: indica las correlaciones nulas con tests que miden aspectos diferentes.

A través de estos dos tipos de validez se podría ir definiendo un constructo psicológico.

2) Análisis Factorial del test

El análisis factorial permite ordenar los datos y facilitar la interpretación de las correlaciones. Se espera un factor explicativo del constructo con saturaciones altas del test y los tests que miden aspectos parecidos, y con saturaciones bajas de aquellos tests que miden aspectos diferentes.

Con frecuencia se habla de la estructura factorial de un test como *validez estructural o validez factorial*

3) Análisis de las diferencias individuales que pone de manifiesto un test

Se refiere al análisis de la distribución de las puntuaciones de test y a comparaciones de estos aspectos en distintas muestras. Diferentes edades, sexos, niveles profesionales, etc. Estas comparaciones no son arbitrarias, sino que se derivan de hipótesis que se hacen en función de los conocimientos que se tiene del constructo.

4) Análisis de los cambios en las diferencias individuales

Se refiere a la investigación diacrónica de los mismos sujetos con el mismo test. Este tipo de estudios permite conocer la estabilidad del rasgo a lo largo del tiempo y a través de situaciones.

5) Análisis lógico de los elementos del test

Se refiere al análisis de ítems del test en relación con el constructo. Aquellos ítems que correlacionan positivamente entre sí, pertenecen al mismo constructo. Esto significa que el análisis de consistencia interna de un test no sólo aporta datos respecto a su fiabilidad, sino a su validez. Si se obtiene un coeficiente de consistencia interna bajo, significa que el test no mide un único constructo.

Todos estos métodos enumerados se complementan entre sí. Se trata de ver el constructo que trata de medir el test desde diferentes ángulos. Por tanto, no existe una única medida de la validez de constructo.

5. CUESTIONES RELATIVAS A LA ESTIMACIÓN DE LA VALIDEZ

1. Valor máximo del coeficiente de validez

El valor máximo que puede alcanzar un coeficiente de validez, estimado mediante la correlación entre el test y el criterio, es menor o igual que su índice de fiabilidad. Cuanto peor medido esté el criterio, o menos fiables sean las puntuaciones obtenidas en el criterio, la prueba de rendimiento, peor va a ser la predicción.

2. Validez y longitud del test

Del mismo modo que la fiabilidad de un test mejora aumentando su longitud, con la validez también ocurre lo mismo. Es decir, cuanto mayor sea el número de ítems, mayor será la validez del test.

3. Validez y variabilidad del grupo

La correlación entre dos variables aumenta conforme lo hace la variabilidad de la muestra. Si restringimos el rango de variabilidad de las puntuaciones de una muestra de sujetos, el coeficiente de validez (la correlación del test con el criterio) será menor de lo que debería.

6. ALGUNAS AMENAZAS A LA VALIDEZ DE LA INVESTIGACIÓN

A lo largo del proceso de investigación se pueden dar una serie de amenazas o sesgos que pueden estar afectando a los resultados, es decir, son variables que pueden convertirse en causas alternativas de los efectos descubiertos. A continuación, se presentan las más destacadas según estemos considerando la validez interna, la validez externa o la validez de constructo (Cook y Campbell, 1979).

VALIDEZ EXTERNA

AMENAZAS	CARACTERÍSTICAS
Interacción selección-tratamiento (validez de población)	Capacidad para generalizar el tratamiento a personas que no pertenezcan al grupo estudiado.
Interacción contexto-tratamiento (validez ecológica)	Capacidad para generalización del tratamiento a situaciones más allá de la estudiada.
Interacción historia-tratamiento (validez histórica)	Capacidad para generalizar el tratamiento a otras ocasiones temporales (pasado o futuro).

VALIDEZ DE CONSTRUCTO

AMENAZAS	CARACTERÍSTICAS
Explicación preoperacional inadecuada	Escasa definición de los constructos.
Sesgo debido al empleo de operacionalizaciones únicas	Medida de una sola variable dependiente.
Sesgo debido al empleo de un método de operacionalización único	Medida de la variable dependiente mediante un solo método.
Adivinación de la hipótesis	Los sujetos adivinan la hipótesis experimental y actúan de la forma que creen que el investigador desea que actúe.
Expectativas del experimentador	Los experimentadores producen sesgos en el estudio a causa de sus expectativas en y durante el estudio.
Confusión entre constructos y niveles de constructo	No se implementan todos los niveles del constructo y pueden presentarse de forma débil o no existir.
Interacción de tratamientos intrasujeto	Los sujetos forman parte también de otros tratamientos.
Interacción de administración de pruebas y tratamientos	La administración de las pruebas puede facilitar o inhibir el efecto del tratamiento.
Generalidad restringida entre constructos	Cuando un constructo no puede ser generalizado de un estudio a otro.

VALIDEZ INTERNA

AMENAZAS	CARACTERÍSTICAS
Historia	Sucesos externos al tratamiento que pueden afectar a la VD.
Maduración	Cambios biológicos y psicológicos de los sujetos que afectarán a sus respuestas.
Administración de pruebas	Efectos del pretest.
Instrumentación	Cambios en la instrumentación o en los observadores.
Selección	Diferencias en los sujetos anteriores al tratamiento.
Regresión estadística	Las puntuaciones extremas tienden a acercarse al postest independientemente del tratamiento.
Mortalidad experimental	Pérdida de sujetos a lo largo del estudio.
Interacciones con selección - ***Selección-maduración*** - ***Selección-instrumentación*** - ***Selección-historia***	Algunas características de los sujetos producen errores en el efecto del tratamiento en el postest; efectos diferenciales en los factores de selección.
Difusión/imitación de los tratamientos - ***Efecto Hawthorne*** - ***Efecto conejo de Indias***	Los miembros de los grupos de tratamiento comparten condiciones de tratamiento con cada uno de los demás o intentan copiar el tratamiento
Igualación compensatoria de tratamientos: ***Efecto de novedad***	Determinar que todos los sujetos, tanto del GE como del GC reciban un tratamiento que les proporcionen efectos beneficiosos.
Frustración de los sujetos: ***Desmoralización***	Los miembros de los grupos que no reciben tratamiento se perciben como inferiores.

BIBLIOGRAFÍA

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

Castejón Costa, J. L. (1997). *Introducción a los métodos y técnicas de investigación y obtención de datos en psicología*. Sant Vicent del Raspeig, España: ECU.

Cook, T. D., y Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin Company.

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

Martínez Arias, M. R. (1995). *Psicometría. Teoría de los tests psicológicos y educativos*. Madrid, España: Síntesis.

Suárez, J. M., Jornet, J. M. y Sáez, A. (1991). *Proceso general de investigación. Validez y diseño.* Documento no publicado. Universidad de Valencia, Valencia, España.

ACTIVIDADES

Identificar las variables y decir de qué tipo son, así como las posibles amenazas a la validez en los siguientes diseños de investigación:

EJEMPLO I:

Se pretende averiguar la eficacia diferencial, en el material de Lenguaje, de tres formas de definición del Dominio Instruccional: facetas de GUTTMAN, CDT de MERRILL y algorítimo de SCANDURA. Dadas las diferencias que pueden existir respecto a la edad de los sujetos, se opta por seleccionar grupos definidos respecto a la misma. A tal efecto se toman, de forma aleatoria, en tres centros piloto de la ciudad de Alicante tres sujetos entre los que cursan cada uno de los tres niveles siguientes: primero de Primaria, segundo de ESO y Bachillerato. Los sujetos reciben la instrucción correspondiente a la primera evaluación de acuerdo con el diseño de facetas de GUTTMAN, los correspondientes a la segunda evaluación de acuerdo con el CDT de MERRILL y los de la tercera según el sistema de SCANDURA. Dos semanas después de la finalización de los contenidos de cada evaluación, su profesor los examina con una prueba mixta (objetiva más desarrollo) y califica su progreso en el apartado correspondiente. Los resultados obtenidos se recogen en la siguiente tabla:

	FACETAS	CDT	ALGORITMICO
1º PRIMARIA	4 5 7	3 6 8	2 7 7
2º ESO	3 4 5	6 2 9	3 5 8
BACHILLERATO	3 3 4	5 4 5	6 7 8

EJEMPLO II:

Se pretende probar la incidencia de la estructuración jerárquica del Dominio Instruccional sobre el rendimiento de los sujetos en Matemáticas a nivel de ESO. Para ello, se seleccionan 6, entre el total de los profesores que se ofrecen como voluntarios para la experiencia, y se les somete a un entrenamiento sobre procedimientos para estructurar jerárquicamente un Dominio Instruccional que se lleva a cabo en los meses estivales. Al comienzo del curso siguiente se administra una prueba de Rendimiento en matemáticas (RM) a los alumnos de estos profesores (Tipo I) y a los alumnos de otros 6 profesores que no han recibido entrenamiento alguno (Tipo II). Transcurrido un semestre se vuelve a administrar una prueba de Rendimiento en Matemáticas a los mismos alumnos mencionados.

Por otra parte, se considera como un factor decisivo en el rendimiento el tipo de actitud que tengan los profesores respecto a la estructuración del Dominio Instruccional para lo cual se les administra una prueba (AC) en la que expresen sus actitudes respecto al principio y fin de la experiencia.

Después de proceder a la selección de un sujeto al azar en cada una de las doce clases, incluidas en la experiencia, se expresan los resultados en una tabla.

EJEMPLO III:

Se pretende determinar la eficacia de un programa para la recuperación de sujetos discalcúlicos, en el momento crítico de 10 años de edad, basado en el desarrollo instruccional de los principios motivacionales de KÉLLER.

Además, se considera relevante para los resultados que se obtengan el nivel de integración que cada sujeto posea en el medio escolar en que se desenvuelve.

Para ello, se seleccionan 60 sujetos de la ciudad de Valencia que presentan esta problemática y que cursan 3º de Primaria y se les administra una prueba sociométrica para determinar su nivel de integración. De acuerdo con los resultados obtenidos, se extraen veinte sujetos repartidos por igual entre los que tienen un nivel de integración elevado (NIA) y los que poseen un nivel deficitario (NIB).

No obstante, antes de dar comienzo la intervención (I) se determina importante obtener información sobre los niveles de Lenguaje de los sujetos implicados en la experiencia. A tal efecto, se elabora una prueba de rendimiento de Lenguaje (L) junto a otro de Cálculo (C) y se administran ambas a todos los sujetos seleccionados.

Posteriormente, los sujetos son sometidos al programa de intervención durante seis meses y se vuelve a recabar información sobre los niveles de Cálculo y de Lenguaje conseguidos hasta ese punto (F), con un oscilación de unos días a dos semanas tras la finalización del programa.