Psicometria: Tests Psicométricos, Confiabilidad y Validez

Jaime Aliaga Tovar

En las ciencias fácticas se miden las características de los objetos que estudian. La psicología es una ciencia fáctica y para medir los atributos o características psicológicas del ser humano utiliza como instrumentos a los tests. Estos pueden ser tests proyectivos o tests psicométricos. Los tests psicométricos son los que utilizan el concepto de medición y tienen su fundamento en la psicometría. El primer test psicométrico fue la Escala Métrica de la Inteligencia creada por los franceses Binet y Simon (1905), con la que se introdujo en psicología el concepto de edad mental. Uniendo este concepto con el de edad cronológica, el psicólogo alemán William Stern (1912) creó el concepto de Cociente Intelectual o CI. Por otro lado, el Cuestionario de Datos de Personalidad de Robert S. Woodsworth (1916) es reconocido como el primer test de personalidad. Desde las primeras décadas del pasado siglo los tests psicométricos han sido construidos siguiéndose el enfoque de la Teoría Clásica de los Tests, creada por el psicólogo inglés Charles Spearman en la segunda década del siglo XX; en las tres últimas décadas del mismo siglo apareció otro enfoque para la construcción de tests que ha sido llamado Teoría de Respuesta al Ítem (TRI), pero todavía hay pocos tests basados en esta teoría. Los tests psicométricos han tenido un gran avance relacionado con los avances de la psicometría que es la rama de la psicología que se ocupa de las mediciones mentales. Tests psicométricos son por ejemplo la Escala de Inteligencia para Adultos de Wechsler (WAIS) y su contraparte para niños (WISC), del mismo autor; otro test de reconocido prestigio es el Test de Matrices Progresivas de Raven; por otro lado, en personalidad, tests psicométricos son, por ejemplo, el Inventario Multifásico de la Personalidad de Minnesotta (MMPI), el

Cuestionario 16PF de Cattell y el Inventario de la Personalidad de Eysenck (EPI).

ENFOQUE ACTUAL

El enfoque actual sobre los tests psicométricos lo haremos empezando por la conceptualización de la psicometría hasta llegar a una breve descripción de la teoría de la respuesta al ítem (TRI), tratando de paso otros conceptos básicos necesarios para comprender la realidad de los tests psicométricos.

Psicometría

Una disciplina de la psicología cuya finalidad intrínseca es la de aportar soluciones al problema de la medida en cualquier proceso de investigación psicológica.

También es un campo metodológico que incluye teorías, métodos y usos de la medición psicológica, en que se incluyen aspectos meramente teóricos y otros de carácter más práctico.

La perspectiva teórica incluye las teorías que tratan de las medidas en psicología, encargándose de describirlas, categorizarlas, evaluar su utilidad y precisión, así como la búsqueda de nuevos métodos, teorías y modelos matemáticos que permitan mejores instrumentos de medida.

La perspectiva práctica se ocupa tanto de aportar instrumentos adecuados para conseguir buenas medidas como de los usos que de los mismos se puedan realizar. Estos instrumentos son los tests psicométricos.

Finalmente, la psicometría se distingue por el uso del lenguaje formal y estructurado de las matemáticas.

Medición

En la psicología, la educación y las ciencias sociales se trata de medir aspectos que no son físicos ni directamente observables. La medición según Nunnally (1987) consiste en reglas para la asignación de números a objetos en tal forma que representen cantidades de atributos. La palabra “objeto” se usa en un sentido amplio e incluye personas. En psicología, medir es dar la magnitud de cierta propiedad o atributo, por ejemplo, la inteligencia, la

extraversión, el razonamiento verbal, de una o más personas, con ayuda del sistema numérico.

Los tests psicométricos son los instrumentos que se utilizan en psicología para la medición de los atributos psicológicos. Es conveniente señalar que:

Test psicométrico

El test psicométrico es un procedimiento estandarizado compuesto por ítemes seleccionados y organizados, concebidos para provocar en el individuo ciertas reacciones registrables; reacciones de toda naturaleza en cuanto a su complejidad, duración, forma, expresión y significado (Rey, 1973).

Requisitos de un test psicométrico

Para que un test sea llamado test psicométrico debe cumplir varios requisitos:

a) El contenido y la dificultad de los ítemes están sistemáticamente controlados (*construcción del test*).

b) La situación de aplicación del test: el ambiente en el cual se le administra, el material del test, la administración, debe estar bien definida y debe ser reproducida idénticamente para todos los sujetos examinados con el test.

c) El registro del comportamiento provocado en el sujeto examinado debe ser preciso y objetivo. Las condiciones de cómo hacer este registro deben estar bien definidas y deben ser cumplidas rigurosamente.

d) El comportamiento registrado debe ser evaluado *estadísticamente* con respecto al de un grupo de individuos llamado grupo de referencia o normativo.

e) Los sujetos examinados son clasificados en función de normas resultantes del examen previo del grupo de referencia o normativo (baremo), lo que permite situar cada una de las respuestas, totales o parciales, en una distribución estadística (*contraste*).

f) Las respuestas a las cuestiones planteadas dan una medida correcta del comportamiento al que el test apunta (*validez*).

g) Si las condiciones no cambian, la repetición del examen debe conducir siempre al mismo resultado, o a otro muy próximo (*fiabilidad*) (Pichot, 1996).

Estandarización

Se llama así al proceso mediante el cual se establecen procedimientos unívocos para la aplicación, calificación e interpretación de un test psicométrico (Cronbach, 1972).

Cuando las condiciones de administración y calificación del test psicométrico están bien definidas y su utilización es idéntica en todos los sujetos examinados, entonces el aspecto más importante que queda por resolver es la interpretación de las puntuaciones logradas por los sujetos evaluados. Esta interpretación se realiza comparando el puntaje obtenido por el sujeto con las puntuaciones contenidas en el baremo o tablas de normas.

Interpretación de los puntajes de un test psicométrico estandarizado

Los números que arrojan la medición de un atributo psicológico realizada con un test psicométrico se denominan puntajes o calificaciones directas. Estos puntajes en sí mismos no tienen un significado preciso, adquieren un significado psicométrico cuando se les compara con una tabla de normas o baremo, que ha sido previamente construida con las puntuaciones que en el test han obtenido un grupo de sujetos llamado grupo normativo. Al realizar esta comparación se puede hacer la clasificación de los sujetos examinados con lo cual se llega a cumplir la finalidad del test, que es clasificar a los sujetos examinados.

Los puntajes directos se transforman en varios tipos de puntajes derivados o unidades de medición que se presentan en las tablas de normas o baremos.

Un baremo es una tabla que sistematiza las normas (afirmación estadística del desempeño del grupo normativo en el test psicométrico) que trasforman los puntajes directos en puntajes derivados que son interpretables estadísticamente. Puntajes derivados son los siguientes: a) percentiles, puntaje derivado que tranforma el puntaje directo en una escala del 1al 100, llamándose cada uno de los puntos un centil. Un examinado que tiene un puntaje

directo equivalente al percentil 80, se dice que supera al 80% del grupo normativo en el aspecto evaluado; b) puntajes estándar, que son aquellos que tienen como unidad a fracciones de la desviación estándar, ejemplos: el puntaje CI, el puntaje eneatipo (escala del 1 al 9), el puntaje decatipo (escala del 1 al 10), entre otros. También es un puntaje derivado la edad mental.

Clasificación de los tests psicométricos

Hay varias clasificaciones. Presentamos la siguiente clasificación:

Criterio	Clasificación
* Por su forma de dar las instrucciones	*Orales *Escritos (aunque en algunos casos hay que ejemplificar lo que se quiere que el sujeto realice en la tarea, como en el caso de personas con problemas auditivos)
* Por su administración	* Individual * Colectiva * Autoadministrada
* Por la forma o tipo de respuesta que exigen	* Objetivas * Subjetivas
* Por el material de la prueba	* Lápiz y papel * Verbal completamente * De ejecución (material, manual, visomotor) * De aparatos especiales * Combinación de los tres primeros (Ej.: WAIS)
*Por su forma de calificación	* Manual * Electrónica
* Por sus edades límites de aplicación	* Pruebas para infantes (baby test) * Pruebas para preescolares * Pruebas para escolares * Pruebas para adolescentes * Pruebas para adultos
*Por su libertad de ejecución	* Pruebas de poder (power test) • Pruebas de rapidez (speedy test)

En los test psicométricos utilizados en educación ha aparecido en los últimos años un tipo de test denominados *Test relacionados con los criterios*, que siendo psicométricos defieren de los otros tests que utilizan el concepto de norma fundamentado en la distribución normal o curva de Gauss. En un test relacionado con el criterio el examinador fija un puntaje que determinara a los aprobados de los desaprobados en un examen. Por ejemplo, puede fijar que de 20 preguntas presentadas será aprobado el alumno que responda correctamente a 18 de las preguntas.

El test psicométrico como auxiliar para una mejor toma de decisiones por parte del psicólogo

En su labor cotidiana el psicólogo debe tomar variadas decisiones, el test psicométrico puede auxiliarle para hacer una mejor para toma de decisiones en los siguientes campos:

- En la **selección**, la decisión consiste, por ejemplo, en aceptar o rechazar a un postulante o solicitante.
- En la **clasificación**, la decisión es tomar el curso alternativo de acción que se debe instigar.
- En el **diagnóstico**, la decisión se hace respecto al tipo de tratamiento pedagógico, psicopedagógico o psicológico a seguir.
- En la **investigación,** se utilizan para tomar decisiones acerca de la elaboración de hipótesis, exactitud en la formulación teórica, recolección de la información necesaria.
- En la **evaluación**, ayudan en la decisión de otorgar, por ejemplo, la calificación que se merece un alumno o el punto hasta el cual un determinado procedimiento es o no eficiente (Brown, 1980).

Limitaciones en el uso de un test psicométrico

Un test psicométrico puede presentar las siguientes limitaciones en su uso:

1) Una prueba o test debe emplearse solamente para apreciar los aspectos para las cuales se ha elaborado.

2) Las normas (baremo) de una prueba no tienen validez universal. Sólo son válidas si los individuos que toman el test poseen características similares a las de los sujetos que formaron la muestra que sirvió para obtener dichas normas.

3) Se deben construir normas para los grupos en los cuales se usará el test, si esos grupos difieren de aquellos en los cuales se hicieron los baremos que aparecen en el manual de la prueba.

4) Un test aprecia la función medida tal como se dan en el individuo en el momento de aplicación de la prueba. Si las condiciones que influyen sobre el individuo se modifican, existe la posibilidad de que tales cambios influyan en los puntajes resultantes del test.

5) Los resultados de un test no deben emplearse para diagnosticar *per se* estados patológicos. Deben considerarse como elementos de información que juiciosamente analizados e integrados con otros elementos de información ayudarán al diagnóstico.

6) Categorías descriptivas (inteligencia muy superior, superior, etc.) sólo deben utilizarse para los tests cuyos puntajes le dieron origen. Por ejemplo, la tabla de categorías del WAIS sólo debe ser utilizada con este test y no con otro (Anastasi, 1982).

Secciones o partes de un test psicométrico

Un test psicométrico tiene generalmente las siguientes secciones o partes:

1. El test propiamente dicho.

2. El manual del test. Documento que contiene los siguientes capítulos o partes:

a) Exposición de los objetivos de la prueba (qué mide). Generalmente empieza con un resumen mostrado en la FICHA TÉCNICA.

b) Descripción de las características estructurales del test (sus partes y componentes).

c) Información acerca del proceso de estandarización o tipificación.

d) Instrucciones generales sobre la manera de aplicar o administrar la prueba y del tipo de población en la cual es aplicable.

e) Descripción del material de examen propiamente dicho a las instrucciones detalladas para la aplicación del test o de cada uno de los subtests.

f) Instrucciones para las valoraciones (calificación) de las respuestas obtenidas en cada uno de los subtests.

g) Información estadística y psicométrica acerca de las propiedades de la prueba como instrumento de medida, vale decir, confiabilidad y validez.

h) Tablas de normas o baremos con los puntajes directos y convertidos para los diferentes grupos de edades y poblaciones (poblaciones de referencia y grupos normativos adecuadamente descritos).

Los puntajes convertidos son, usualmente, los percentiles y los puntajes estándar (Cronbach, 1972).

Cualidades que debe tener un test psicométrico

Confiabilidad

La confiabilidad (o consistencia) de un test es la precisión con que el test mide lo que mide, en una población determinada y en las condiciones normales de aplicación. (Anastasi, 1982; Aiken, 1995). (Las condiciones normales de aplicación se refieren a las condiciones especificadas en el manual del test).

La falta de confiabilidad de un test psicométrico esta en relación con la intervención del error. Se considera que el error es cualquier efecto irrelevante para los fines o resultados de la medición que influye sobre la falta de confiabilidad de tal medición. El error es de dos tipos: a) Error constante (sistemático), que se produce cuando las mediciones que se obtienen con una escala son sistemáticamente mayores o menores que lo que realmente deben ser. b) Error causal (al azar o no sistemático), que se produce cuando las medidas son alternativamente mayores o menores de lo que realmente deben ser. Este último tipo de error interviene cuando se afecta la confiabilidad de un test psicométrico. Este error tiene que ver con la salud, fatiga, motivación, tensión emocional, fluctuaciones de la memoria, condiciones externas de luz, humedad, ventilación, calor, distracción por problemas del momento, familiaridad con la prueba, que presenta el examinado al momento de dar el test (Rey, 1972; Brown, 1982).

¿Cómo se presenta la confiabilidad de un test psicométrico? La confiabilidad se presenta por medio del coeficiente de confiabilidad (r_{xx}) y del error estándar de medida (EEM).

A) Coeficiente de confiabilidad

Es un coeficiente de correlación entre dos grupos de puntajes e indica el grado en que los individuos mantienen sus posiciones dentro de un grupo. Abarca valores desde 0 a 1. Cuanto más se acerque el coeficiente a 1, más confiable será la prueba.

El coeficiente de confiabilidad señala la cuantía en que las medidas del test están libres de errores casuales o no sistemáticos. Por ejemplo, un coeficiente de 0.95 quiere decir que en la muestra y condiciones fijadas de aplicación del test el 95% de la varianza de los puntajes directos se debe a la auténtica medida, y sólo el 5%, a errores aleatorios.

Existen cuatro métodos básicos para obtener el coeficiente de confiabilidad (r_{xx}): Método de las formas equivalentes; método del test-retest; método de la división por mitades emparejadas o "split half method"; y método de la equivalencia racional o de Kuder-Richardson.

a) Método de las formas equivalentes: Se aplican dos formas equivalentes o paralelas del test al mismo grupo de individuos, y las dos series de puntajes resultantes se correlacionan con el coeficiente producto de los momentos de Pearson (r).

b) Método del test-retest: Se aplica dos veces el mismo test (el lapso entre las aplicaciones se determina previamente), a una misma muestra de individuos. Las dos series de puntajes resultantes se correlacionan con el coeficiente de correlación "r" de Pearson.

c) Método de la división por mitades emparejadas o "split half method": Se aplica el test una sola vez a una muestra. Luego, se califica por separado los ítemes pares (2, 4, 6, ..., n) y los ítemes impares (1, 3, 5, ..., n). A continuación, las dos series de puntajes resultantes se correlacionan con el coeficiente "r" de Pearson, pero por haberse dividido el test en dos partes (ítemes pares e ítemes impares), el "r" resultante debe ser "corregido" para arrojar el "r" para todo el test. Esta corrección se efectúa con la fórmula de profecía de Spearman-Brown:

d) Método de la equivalencia racional: En este método se considera que si un test esta formado por un conjunto de ítemes estos pueden ser considerados como un conjunto de tests paralelos (tantos como ítemes tenga el test). Luego se deriva una ecuación para computar el coeficiente de confiabilidad. Kuder y Richardson derivaron varias fórmulas para el cálculo del coeficiente de

confiabilidad, son las más conocidas la KR_{20} y la KR_{21}. Actualmente, un coeficiente más utilizado es el coeficiente alfa de Cronbach (1972; Anastasi, 1982; Aiken, 1995).

B) Error estándar de medida

Por medio de este error estándar de medida se estima el intervalo probable de puntajes en el cual se encontrará el puntaje verdadero de un sujeto examinado con un test psicométrico.

El error estándar de medida (EEM) se obtiene a través de la siguiente fórmula:

$$EEM = s\sqrt{1-r_{xx}}$$

Donde:

s = Desviación estándar de los puntajes de la distribución.

r_{xx} = Coeficiente de confiabilidad del test.

1 = Constante.

Obtenido el EEM, debemos escoger el nivel de confianza:

- Nivel de confianza del 68% = PD ± 1 EEM.
- Nivel de confianza del 95% = PD ± 2 EEM.

Para el nivel de confianza del 68% la interpretación es la siguiente: "Podemos concluir, con un 68% de confianza, que el puntaje verdadero de un sujeto está en la zona o intervalo comprendido entre su puntaje directo u obtenido (PD) y ± 1 EEM".

El nivel de confianza más usado en psicometría es el del 95%: "el puntaje verdadero de un sujeto se encontrara en el intervalo comprendido entre su puntaje obtenido o directo (PD) y ± 2 EEM".

Validez

Si tenemos una prueba “X” nos equivocaríamos al creer que su titulo nos dice lo que la prueba mide, pues cualquier persona puede reunir un conjunto de reactivos y esperar a obtener una medida, por ejemplo, de razonamiento numérico o de las estrategias de aprendizaje. La averiguación de lo que la prueba mide no responde a la pregunta ¿cómo llama el autor a la prueba?, sino más bien ¿a qué hacen referencia los puntajes obtenidos en esta?, ¿es válido el uso o la interpretación de las puntuaciones de este test?, ¿qué generalizaciones se pueden hacer apropiadamente a partir de la puntuación en esta prueba? (Thorndike, 1989). En esencia, el trasfondo de estas preguntas es determinar cuáles son los procesos mentales que pone en juego el test. Ahora bien, el responder a las

citadas interrogantes necesita de una indagación larga y compleja que en psicometría se denomina *proceso de validación.*

a) **Distinción entre la validez y la confiabilidad según el error.** La distinción entre confiabilidad y validez se basa en lo que consideramos como error. En la validez interesan los errores constantes o sistemáticos y en la confiabilidad los errores aleatorios o no sistemáticos. El siguiente ejemplo nos permitirá precisar la diferencia entre ambos tipos de error: Supongamos que un reloj es adelantado 20 minutos. Si se trata de un buen cronometro el tiempo que marca será confiable (es decir consecuente), pero no será valido en comparación con el tiempo estándar (hora GMT).

b) Definición de validez. *En términos estadísticos* la validez se define como la proporción de la varianza verdadera que es relevante para los fines del examen. Con el término relevante nos referimos a lo que es atribuible a la variable, características o dimensión que mide la prueba.

En este sentido, generalmente la validez de un test se define ya sea por medio de (1) la relación entre sus puntuaciones con alguna medida de criterio externo, o bien (2) la extensión con la que la prueba mide un rasgo subyacente específico hipotético o "constructo".

*En términos psicométricos***,** la validez es un concepto que ha pasado por un largo proceso evolutivo, desde aquella posición que sostenía que "un test es válido para aquello con lo que correlaciona" (Guilford, 1946, citado en Muñiz, 1996, p. 52), hasta la más reciente que la entiende como un juicio evaluativo global en que la evidencia empírica y los supuestos teóricos respaldan la suficiencia y lo apropiado de las interpretaciones y acciones en base a los puntajes de las pruebas, que son función no sólo de los ítemes sino también de la forma de responder de las personas así como del contexto de la evaluación.

Es decir, lo que se valida no es la prueba sino las inferencias hechas a partir de la misma, lo que tiene dos importantes consecuencias: a) el responsable de la validez de una prueba ya no es solo su constructor sino también el usuario, y b) la validez de una prueba no se establece de una vez por todas sino que es resultado del acopio de evidencias y supuestos teóricos que se dan en un proceso evolutivo y continuo que comprende todas las cuestiones

experimentales, estadísticas y filosóficas por medio de las cuales se evalúan las hipótesis y teorías científicas (Messick, 1995).

En este contexto, el concepto *validez* refiere a la adecuación, significado y utilidad de las inferencias específicas hechas con las puntuaciones del los tests. La validación de un test es el proceso de acumular evidencia para apoyar tales inferencias. Una variedad de evidencias pueden obtenerse de las puntuaciones producidas por un test dado, y hay muchas formas de acumular evidencia para apoyar una inferencia específica. La validez, sin embargo, es un proceso unitario. Aunque la evidencia puede ser acumulada de muchas formas, la validez se refiere siempre al grado en que esa evidencia apoya las inferencias que se hacen a partir de las puntuaciones" (APA, AERA, NCME, 1985, citado en Gómez e Hidalgo, 2002, p. 2). La validez no se puede resumir en un solo indicador o índice numérico, al igual que ocurre con la confiabilidad (p.e., el coeficiente de confiabilidad), sino que la validez de las puntuaciones de un test se asegura mediante la acumulación de *evidencia* teórica, estadística, empírica y conceptual del uso de las puntuaciones.

c) Tipos de evidencia. En 1954 un comité presidido por L. J. Cronbach estableció por encargo de la Asociación de Psicología Americana (APA), que la validez era de cuatro tipos: validez de contenido, validez predictiva, validez concurrente y validez de constructo. Actualmente se coincide, desde el punto de vista científico, que la única validez admisible es la validez de constructo (Messick, 1995). Validación que ha de hacerse en un marco teórico, pues se trata en última instancia de confirmar o explicar las inferencias que se hagan de los puntajes.

La validez de constructo esta referida al grado en que cada prueba refleja el constructo que dice medir, elaborándose operativamente cuando el usuario desea hacer inferencias acerca de conductas o atributos que pueden agruparse bajo la etiqueta de un constructo particular. Su lógica en muchos aspectos así como en sus métodos, es esencialmente la del método científico, pudiendo verse como la elaboración de una "miniteoría" acerca de una prueba (Kline, 1985) cuyas hipótesis deben someterse a contraste con evidencias que provengan de diferentes fuentes como la de los tipos de validez propuestos por Cronbach, entendidas como estrategias de validación, en vista que cada tipo de inferencia requerirá una

estrategia distinta para la obtención de las evidencias (Vidal, 1996, en Muñiz, 1996).

En el estudio de la validez de constructo estas evidencias están relacionadas a cinco aspectos: a**) Contenido** (relevancia y representatividad del test); b) **Sustantivo** (razones teóricas de la consistencia observada de las respuestas); c) **Estructural** (configuración interna del test y *dimensionalidad*); d) **Generalización** (grado en que las inferencias hechas a partir del test se pueden generalizar a otras poblaciones, situaciones o tareas); e) **Externo** (relaciones del test con otros tests y constructos); f) **Consecuencia** (consecuencias éticas y sociales del test) (Messick, 1995).

d Categorías de la validez. La validez empieza a considerarse como el grado en que cada test refleja el constructo que dice medir y que las relaciones entre tests que miden distintos constructos reflejan las relaciones hipotetizadas entre ellos. En este sentido, al estimarse que la validez de un test es la validez de constructo la que ha de hacerse en un marco teórico, ya no se tiende a hablar de tipos de validez sino de categorías o estrategias de validación comprendiendo éstas a los tipos tradicionales de validez: validez de contenido, validez empírica y validez de constructo. Si tenemos en cuenta que lo que se valida no es el test sino las inferencias hechas a partir del mismo, cada tipo de inferencia requerirá una estrategia distinta. (Vidal, 1996, en Muñiz, 1996).

(**1) Validez de Contenido (evidencia del contenido).-** ¿Los ítems que constituyen el test son realmente una muestra representativa del ***dominio*** de contenido o dominio conductual que nos interesa?

Es conveniente precisar que un ***dominio*** o campo conductual es una agrupación hipotética de todos los reactivos posibles que cubren un área psicológica particular. Al hablar de este conjunto de reactivos posibles, se emplean los términos de dominio, universo o población conductual como sinónimos. Por ejemplo: Un test de vocabulario debe ser una muestra adecuada del dominio o universo de ítems posibles en esta área.

La validez de contenido consiste en determinar lo adecuado del muestreo de reactivos del universo de reactivos posibles; en este sentido, es una "medida" de lo adecuado del muestreo. Ponemos "medida" entre comillas debido a que este tipo de validez consiste en

una serie de estimaciones u opiniones, que no proporcionan un índice cuantitativo de validez (para su obtención no se utiliza procedimientos estadísticos). Este tipo de validez se asocia fundamentalmente a los tests de aprovechamiento o rendimiento (test de matemática, historia, etcétera); aunque no existen razones para que no pueda aplicarse a los otros tipos de pruebas psicológicas (pruebas de aptitudes, habilidades, etcétera).

Para su determinación se compara sistemáticamente los reactivos del test con el dominio conductual del contenido postulado. Por ejemplo: si tenemos una lista de 500 palabras que esperamos que los estudiantes de un curso sean capaces de escribirlas correctamente al final de este, su *performance* o rendimiento respecto a estas palabras será importante solamente en tanto que proporciona una prueba de su habilidad para escribir correctamente las 500 palabras. El test que construyamos tendrá una muestra de las 500 palabras, pero sólo tendrá validez de contenido en la medida en que proporcione una muestra adecuada de las 500 palabras que represente. Si seleccionamos solamente palabras fáciles o difíciles, o palabras que representen únicamente ciertos tipos de faltas comunes de ortografía, estaríamos propensos a obtener una validez de contenido muy baja. En consecuencia, el aspecto clave en la validez de contenido es el muestreo de los reactivos. En otras palabras, la validez de contenido es cuestión de determinar si la muestra de sus reactivos es representativa del universo o dominio conductual de ítems al que supuestamente representa.

Para hacer esta determinación se recurre a "jueces" (o expertos, generalmente en número impar). El proceso es básicamente lógico y racional, los distintos jueces pueden no estar de acuerdo en la validez de contenido de un test; por ejemplo, la falta de claridad en la especificidad del dominio conductual, hará que resulten difíciles los juicios de validez de contenido. Existen algunos índices estadísticos para valorar el grado de acuerdo de los jueces en torno a los reactivos, por ejemplo el coeficiente V de Aiken.

Un procedimiento para que el proceso de "enjuiciamiento" de los reactivos sea lo más objetivo posible, es el siguiente:

- El constructor de la prueba:

* Define específicamente el dominio del contenido por medio de una descripción que lo debe delimitar claramente.

* Define, si fuera necesario, subcategorías importantes del dominio, especificando esta importancia en términos porcentuales.

- Los jueces:

* Determinan si los reactivos sometidos a su consideración pertenecen o no al dominio definido así como también si, tomados en conjunto, tienen una proporción adecuada.

* También enjuician la bondad de la redacción de los elementos.

Es usual considerar en los tests de aprovechamiento escolar a este tipo de validez como un concepto similar al de validez curricular. Por otro lado, es necesario diferenciar la validez de contenido de la llamada validez *de facie*. Esta última se da cuando se revisa superficialmente los reactivos y se consideran que los ítems "parece" que miden lo que se supone tienen que medir. Esta validez puede ser una consideración importante a tener en cuenta, si la "apariencia" de los ítems influye en la motivación del sujeto. Por ejemplo, si en un test para adultos se incluyen reactivos en lenguaje y contenido infantil, se dirá que este test no tiene validez *de facie*; el sujeto puede no sentirse motivado a obtener buenos resultados al sentir que la prueba es poco importante para la decisión que se va a tomar.

(2) Validez Predictiva (evidencia externa) - ¿Predicen las puntuaciones del test un rendimiento o conducta futura? (Junto con la validez concurrente se le denomina también validez empírica del test).

Un uso común de los tests es predecir la conducta futura; utilizamos el test para ayudarnos a tomar alguna decisión práctica (selección, clasificación, etc.). En cada una de estas situaciones, cuanto mayor es la exactitud de predicción del resultado (es decir del criterio externo), tanto más útil será la prueba. Por ejemplo, el test será un componente aceptable de un proceso de selección de personal, si sus calificaciones o puntuaciones predicen la ejecución de algún componente importante del trabajo (criterio externo); en otras palabras, para que el test se pueda utilizar como parte de un proceso de selección es preciso demostrar la validez de la prueba relacionándola con los criterios pertinentes. En este sentido, el contenido de la prueba pasa a tener un lugar secundario, siendo el interés fundamental del psicólogo el averiguar si el test predice un criterio determinado.

Para este logro es necesario que los criterios externos con los cuales se relacionará las puntuaciones del test sean criterios validos y confiables.

Un *criterio* es cualquier desempeño que los sujetos tienen en la vida real, por ejemplo, las medidas de rendimiento académico, medidas de rendimiento laboral, clasificaciones psiquiátricas, etcétera. En muchos casos resulta imposible hallar un *criterio* no ambiguo de un rasgo mental. Por ejemplo, dos psicólogos, Carla y Abel, que investigan el rasgo de aptitud numérica pueden emplear diferentes criterios externos para correlacionar los puntajes del test que han creado. Así, Carla puede considerar que el criterio externo más adecuado son las calificaciones que reciben los sujetos en un curso de mecánica en taller; mientras que Abel puede considerar como criterio el periodo de tiempo que gastan los estudiantes en aprender una tarea mecánica y sencilla durante el entrenamiento en un fabrica. ¿Qué sucede si las pruebas que emplean ambos psicólogos correlacionan 0.006 con uno de los criterios, y 0.70 con el otro?, ¿cómo podemos afirmar que la prueba es valida cuando arrojan resultados de cierta clase?, ¿se trata en verdad de una prueba de aptitud mecánica? En razón a situaciones como esta se llego a la conclusión de que la validación de un test es un proceso largo y no un hecho aislado. Solamente a través de estudios de correlación con una amplia variedad de criterios podremos comprender que mide la prueba. Así, una serie de investigaciones sobre la "prueba de actitud mecánica" nos puede demostrar que en realidad esta midiendo la habilidad para realizar movimientos fino y cuidadosamente controlados, siendo completamente independiente para comprender las reacciones complejas de la piezas mecánicas. De esta manera el test puede tener una alta correlación con las calificaciones obtenidas en el taller y ninguna con los trabajos e maquinarias.

En el proceso de validación, la validez predictiva de un test (y también la concurrente) se expresa generalmente por medio de un coeficiente de correlación entre los puntajes y los denominados criterios. Este coeficiente se denomina *coeficiente de validación***.** La interpretación de este coeficiente requiere un dominio excelente del análisis estadístico utilizado para obtenerlo. Después del criterio, los procedimientos estadísticos adquieren vital importancia para obtener esta categoría de validez. Incluso para un mejor análisis es conveniente contar con el dispersigrama o scatergrama o "nube de

puntos" (gráfica del coeficiente de correlación entre las puntuaciones del test "X" y las del criterio "Y").

(3) Validez Concurrente.- ¿Permiten las puntuaciones del test la valoración de ciertos hechos presentes? Para estimarla se administra el test y se le correlaciona con el criterio. La diferencia con la validez predictiva se da en dos aspectos: a) las medidas del test y del criterio son obtenidas contemporáneamente, y b) en su uso principal. Respecto a esto último, se la utiliza principalmente para obtener tests como sustitutos de otros procedimientos menos convenientes por diversas razones. Ejemplos: un test de inteligencia colectiva se compara con uno de inteligencia individual. Los diagnósticos de lesiones cerebrales basados en el test de diseños de bloques (cubos de Kohs) se comparan con síntomas neurológicos.

El problema principal de este tipo de validez es encontrar tests que sirvan como criterios válidos y confiables. Análogamente a la validez predictiva requiere un amplio domino de las técnicas de correlación y de los procedimientos estadísticos que se utilizan en su obtención. Junto con la validez predictiva es importante en ciertos problemas de psicología aplicada como en psicología clínica, psicología educacional, psicología industrial y en general, en la toma de decisiones que debe hacer el psicólogo en situaciones de selección, clasificación, hospitalización, etc.

(4) Validez de Constructo.- El constructo viene a ser un concepto hipotético que forma parte de las teorías que intentan explicar la conducta humana: inteligencia, creatividad, dependencia de campo, etc. La validez de constructo es la obtención de evidencias que apoyan que las conductas observadas en un test son (algunos) indicadores del constructo. Este tipo de validez responde a la pregunta "¿cómo se puede explicar psicológicamente la puntuación del test?". La respuesta a esta pregunta puede verse como la elaboración de una "miniteoría" acerca de una prueba psicológica. La lógica de la validez de constructo en muchos aspectos así como en sus métodos, es esencialmente la del método científico.

El proceso de validación de constructo implica a partir del establecimiento de deducciones de la teoría:

a) Formular hipótesis y relaciones entre elementos del constructo, de éste con otros constructos de la teoría y con otros constructos externos.

b) Seleccionar ítemes o tests (indicadores) que representen manifestaciones concretas del constructo.

c) Recogida de datos.

d)Establecer consistencia entre datos e hipótesis, y examinar el grado en que los datos podrían explicarse mediante hipótesis alternativas.

Hay diversos procedimientos para establecer la validez de constructo. Si elaboramos una "miniteoría" esta tendrá tres pasos: (1) en base a la teoría sostenida en ese momento respecto del test, el psicólogo deduce ciertas hipótesis sobre la conducta esperada de las personas que obtienen puntajes diferentes en el test, (2) se reúne datos que confirman o no esas hipótesis, (3) en base a los datos acumulados, se toma la decisión relativa a si la teoría explica adecuadamente los datos. Si no es así se tiene que revisar la teoría y repetir el proceso hasta lograr una explicación más adecuada. El proceso de validación, en ese sentido, es de continua reformulación y refinamiento.

Al determinar la validez de construcción, el propósito es identificar todos los factores que influyen en la ejecución del test y determinar el grado que influyen cada uno de ellos.

Ejemplo: Un psicólogo construye un test de ansiedad y elabora una "microteoría" cuya contrastación le dirá si el test tiene validez de constructo. Las hipótesis a verificar son las siguientes:

1. Los que obtienen puntuaciones elevadas acabarán probablemente en clínicas psiquiátricas en comparación con aquellos de puntuaciones más bajas.
2. Será más fácil que les receten drogas psicotrópicas a los que tienen altas puntuaciones que a los de bajos puntajes.
3. Los hijos de los de puntuaciones altas tendrán mayores probabilidades de tener una puntuación alta en el test que los hijos de quienes tuvieron puntuaciones bajas.
4. El test de ansiedad se correlacionará alta y significativamente (más allá de 0.60) con otros test de ansiedad.
5. El test de ansiedad no se correlacionará con variables que no resulten conexas con la misma.
6. Los grupos psiquiátricos caracterizados como ansiosos alcanzarán en el test unas puntuaciones más altas que los del grupo control.

7.En el test de ansiedad, los sujetos evaluados por supervisores y colegas como ansiosos, lograrán mayores puntuaciones que quienes están considerados como no ansiosos (Kline, 1985).

Los resultados de los estudios que hagamos realmente no "validan" o "prueban" la teoría completa, puesto que nunca se puede demostrar una "construcción" en forma absoluta; solamente se puede aceptar como la mejor definición de trabajo. Si los resultados son negativos, hay por lo menos tres interpretaciones posibles: a) la prueba puede no medir el "constructo", b) el marco teórico puede ser erróneo, permitiendo que se hicieran inferencias incorrectas, o bien c) quizá, el diseño del experimento no permitía una prueba apropiada de la hipótesis. La falla del diseño suele ser la falla más fácil de detectar, pero no siempre se puede hallar con facilidad el lugar exacto de la falla. La interpretación ambigua de los resultados negativos es un inconveniente evidente del procedimiento de validación de los "constructos" (Crombach, 1972; Kline, 1985).

e) Implicaciones prácticas en la validación de un test. El psicólogo que utiliza un test debe tener en cuenta lo siguiente: a) antes de tomar de decisiones sobre individuos o grupos, debe acumular toda la información disponible acerca del test; b) para la predicción o selección, el test debe estar validado en la situación específica donde se va utilizar; c) en cualquier situación, el psicólogo debe tener presente que nuestras ideas sobre la naturaleza de los rasgos y sobre todo lo que miden se modifica constantemente con nueva información hacer

MÉTODOS EMPLEADOS PARA ESTIMAR LA VALIDEZ DE CONSTRUCCIÓN

a) Métodos intrapruebas, cuyas fuentes de evidencia más usadas son: la validez de contenido de la prueba, el análisis de los procesos psicológicos empleados al responderla (p. e., pidiendo que los sujetos "razonen" en voz alta sus respuestas). Otras técnicas estudian la estructura interna de la prueba, mediante el análisis de los ítemes y las correlaciones entre los diferentes subtests; asimismo, también mediante el establecimiento de la homogeneidad a través del coeficiente alfa de Cronbach o los coeficientes de Kuder-Richardson (que contribuyen a evaluar la unidimensionalidad del test).

b) Métodos interpruebas: Utilizan las técnicas del análisis factorial (para evaluar los factores que subyacen en las intercorrelaciones de las

pruebas estudiadas), la validez congruente (en tanto correlaciona los puntajes de la prueba con los puntajes obtenidos en otra prueba de validez ya establecida), los estudios de validez convergente y divergente-discriminante (propuestos por Campbell).

c) El método de los estudios relacionados con los criterios: Que implican la diferenciación de grupos (evaluando la capacidad de la prueba para poder diferenciar dos o más grupos naturalmente separados o diseñados experimentalmente) y los coeficientes de validez (cuando la prueba es aplicada a un grupo de sujetos en los que se estudian criterios relacionados con el constructo teórico estudiado).

d) El método de la manipulación experimental: Se manipula experimentalmente una variable y se observa sus efectos sobre los puntajes de una prueba psicológica o la relación de estos puntajes con algún criterio.

e) El método de los estudios de la capacidad de generalización: Estos estudios analizan sistemáticamente la prueba psicológica en una amplia gama de dimensiones o en condiciones diferentes de administración (p. e., la matriz multirasgo-multimétodo propuesta por Campbell y Fiske).

Teoría de respuesta al ítem (TRI)

Llamada también *Teoría del Rasgo Latente*, es un modelo probabilístico que permite conocer la información proporcionada por cada ítem, y así crear tests individualizados, es decir, a medida. Es un modelo complejo que se ha popularizado como modelo de Rasch (1980) (aunque específicamente el modelo de Rasch es un parámetro de la dificultad del ítem), pero existe también el modelo de dos parámetros, que tiene en cuenta también la discriminación o pendiente de la curva, y el de tres parámetros que tiene en cuenta el factor azar en la respuesta a ítemes de alternativas múltiples (Cortada de Kohan, 1999). La diferencia principal entre el modelo de la Teoría Clásica de los Test y este modelo es que la relación entre el puntaje observado y el rasgo o la aptitud en la teoría clásica es de tipo lineal (PD = PV + e: puntaje directo del sujeto es igual a su puntaje verdadero más el error); mientras que en los diversos modelos de la TRI las relaciones son funciones de tipo exponencial, principalmente logísticos.

Los postulados básicos de la TRI son:

1) El resultado de un examinado en un ítem puede ser explicado por un conjunto de factores llamados rasgos o aptitudes simbolizados por θ.

2) La relación entre la respuesta a un ítem y el rasgo latente se describe como una función monotónica creciente que es la curva característica del ítem.

3) En la TRI los parámetros de aptitud y de los ítemes son invariantes.

Los supuestos de la TRI son:

1) La *unidimensionalidad* del rasgo latente, es decir, que los ítemes de un test deben medir una sola aptitud o rasgo; y

2) La *independencia*, es decir, que las respuestas de un examinado a cualquier par de ítemes son independientes.

Para estimar los parámetros de la TRI se usa el método de *máxima verosimilitud*, que es un proceso complejo que se logra con los *softwares* apropiados como BILOG, BICAL, y otros. (Cortada de Kohan, 1998).

CONCLUSIONES

1) La psicometría es una disciplina de la psicología cuya finalidad intrínseca es la de aportar soluciones al problema de la medida en cualquier proceso de investigación psicológica; constituye, por ello, un campo metodológico que incluye teorías, métodos y usos de la medición psicológica, tanto a nivel teórico como a nivel práctico.

2) En psicología, medir es dar la magnitud de cierta propiedad o atributo, por ejemplo, la inteligencia, la extraversión, el razonamiento verbal, de una o más personas, con ayuda del sistema numérico.

3) El test psicométrico es un procedimiento estandarizado compuesto por ítemes seleccionados y organizados, concebidos para provocar en el individuo ciertas reacciones registrables; reacciones de toda naturaleza en cuanto a su complejidad, duración, forma, expresión y significado.

4) Los requisitos de un test psicométrico son: Construcción del test; la situación de aplicación del test; el registro del comportamiento provocado en el sujeto examinado, que debe ser preciso y objetivo; el comportamiento registrado evaluado estadísticamente con respecto a un grupo de individuos denominado grupo de referencia o grupo normativo; clasificación de los sujetos examinados en función de normas resultantes del examen previo del grupo de referencia o normativo (baremo), lo que permite situar cada

una de las respuestas, en una distribución estadística (contraste);.las respuestas a las cuestiones planteadas deben dar una medida correcta del comportamiento al que el test apunta (validez); y por último, la repetición del examen debe conducir siempre al mismo resultado (fiabilidad).

5) La estandarización es el proceso mediante el cual se establecen procedimientos unívocos para la aplicación, calificación e interpretación de un test psicométrico.

6) La interpretación de los puntajes de un test psicométrico estandarizado se refiere al significado que se les da a los puntajes obtenidos por un grupo de sujetos, al compararlos con una tabla de normas o baremo, estableciendo una clasificación de acuerdo a la conversión de sus puntajes directos en puntajes percentiles, eneatipos o decatipos.

7) Los tests psicométricos se clasifican por lo siguiente: Por su forma de dar las instrucciones; por su forma de administración; por la forma o tipo de respuesta que exigen; por el material de la prueba; por su forma de calificación; por sus edades límites de aplicación; y por su libertad de ejecución. En los test psicométricos utilizados en educación existe actualmente un tipo de tests denominado test relacionado con el criterio.

8) El test psicométrico es empleado como instrumento auxiliar para una mejor toma de decisiones: En la selección, clasificación, diagnóstico, investigación, y evaluación de un determinado grupo de personas.

9) Entre las limitaciones en el uso de un test psicométrico tenemos: Una prueba o test sólo puede medir aquellos aspectos para los que ha sido construido; las normas (baremo) de una prueba no tienen validez universal; si las condiciones que influyen sobre el individuo se modifican, existe la posibilidad de que tales cambios durante la aplicación de la prueba influyan en los puntajes resultantes del test; los resultados de un test no deben emplearse para diagnosticar *per se* estados patológicos;.y, finalmente, las categorías descriptivas (inteligencia muy superior, superior, etc.) sólo deben utilizarse para los tests cuyos puntajes le dieron origen.

10) Un test psicométrico tiene generalmente las siguientes secciones o partes: El test propiamente dicho, y el manual del test.

11) Entre las cualidades que debe tener un test psicométrico, hay que tener en cuenta: la confiabilidad, que puede estimarse a

través de dos procedimientos: el coeficiente de confiabilidad y el error estándar de medida (EEM); y la validez, que es el grado en que una prueba mide lo que intenta medir.

12) La validez científica de un test la da la validez de constructo. La tendencia ya no es hablar de tipo de validez, sino de categorías de validez en la que la validez de contenido, validez predictiva, validez concurrente y validez de constructo reconstituyen en estrategias de validación.

13) La teoría de respuesta al ítem (TRI), denominada también *Teoría del Rasgo Latente*, es un modelo probabilístico que permite conocer la información proporcionada por cada ítem, y así crear tests individualizados, es decir, a medida. Los supuestos de la TRI son: la unidimensionalidad del rasgo latente; y la independencia, es decir, que las respuestas de un examinado a cualquier par de ítemes son independientes.

BIBLIOGRAFÍA

Aiken, L. (1996). *Tests psicológicos de evaluación*. México: Prentice-Hall.

Anastasi, A. (1986). *Los tests psicológicos*. Madrid: Aguilar.

Adkins, D. (1994). *Elaboración de tests. Desarrollo e interpretación de los tests de aprovachamiento*. México: Trillas.

Cerdá, E. (1984). *Psicometría general*. Barcelona: Herder.

Monroe Miller, D. (1974). *Resultados de pruebas psicológicas*. Interpretación estadística. México: Limusa.

Ebel, R. (1977). *Fundamentos de la medición educacional*. Buenos Aires: Guadalupe.

Brown, G. F. (1980). *Principios de la medición en psicología y educación*. México: El Manual Moderno.

Cortada de Kohan, N. (1999). *Teorías psicométricas y construcción de tests*. Buenos Aires: Lugar.

Cronbach, L. J. (1972). *Fundamentos de la exploración psicológica*. Madrid: Biblioteca Nueva.

Gronlund, N. (1978). *La elaboración de tests de aprovechamiento*. México: Trillas.

Kerlinger, F. (1975). *Investigación del comportamiento. Técnica y metodología*. México: Interamericana.

Levine, Ch. y Freeman, F. (1973). *Introducción a la medición en psicología y educación*. Buenos Aires,: Paidós.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Erlbaum.

Magnusson, D. (1969). Teoría de los tests. México DF, México: Trillas.
Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. Educational measurement: Issues and Practice, 14, 5-8.
Morales, M. L. (1996). *Psicometría aplicada*. México: Trillas.
Muñiz, J. (Coord.) (1996). *Psicometría*. Madrid: Universitas.
Muñiz, J. (1990). *Teoría de respuesta a los ítems*. Madrid: Pirámide.
Muñiz, J. (1994). *Teoría clásica de los tests*. Madrid: Pirámide.
Nunnally, J. y Bernstein, Y. (1995). Teoría psicométrica. México: McGraw-Hill.
Tyler, L. (1972). Pruebas y medición en psicología. Madrid,: Prentice-Hall International.