TEMA 3: LOS INSTRUMENTOS DE EVALUACIÓN PSICOLÓGICA

TEMA 3: **LOS INSTRUMENTOS DE EVALUACIÓN PSICOLÓGICA**

INTRODUCCIÓN

En la **evaluación psicológica** el psicólogo realiza una recopilación e integración de datos que obtiene principalmente de "instrumentos" con el fin de realizar un diagnóstico, orientación, selección... Y la **prueba** o**instrumento de evaluación psicológica** mide las variables psicológicas a través de procedimientos diseñados para obtener una muestra de comportamiento.

La formación del buen profesional de la psicología debe contemplar conocer el proceso de evaluación y en el transcurso del mismo **saber elegir y aplicar los instrumentos que le permitirán realizar una labor de calidade interpretar las puntuaciones obtenidas**:

- Definir qué es un instrumento de evaluación psicológica.
- Estar familiarizado con las principales clasificaciones de los instrumentos de evaluación psicológica.
- Saber qué profesionales pueden aplicar los instrumentos de evaluación psicológica.
- Aprender a elegir la mejor prueba de evaluación.
- Identificar las partes de un manual de aplicación de una prueba y su correcta utilización.
- Saber cuáles son los criterios de calidad exigibles a cualquier instrumento de evaluación, que garantizarán los resultados de nuestra labor profesional.
- Conocer las principales unidades de medida de los tests.

1. ¿QUÉ ES UN INSTRUMENTO DE EVALUACIÓN PSICOLÓGICA?

Según García-Montalvo un instrumento de evaluación psicológica es "todo aquello que el evaluador puede utilizar como legítima fuente de datos acerca de un sujeto." En este sentido, un instrumento puede equiparase al vocablo "técnica". Desde nuestro punto de vista se trata de un concepto más general puesto, que los instrumentos de evaluación psicológica se clasifican en función del distinto tipo de **técnicas**, entendidas como conjunto de procedimientos y recursos de que se sirve una ciencia o un arte.

Muchas veces se utiliza indistintamente instrumento, técnica y test, pero no son iguales. Los**instrumentos o técnicas** pueden ser procedimientos no cuantificados ni tipificados como p. ej la entrevista. Los**tests** son un instrumento sistemático y tipificado que compara la conducta de dos o más personas. Sin embargo, a cualquier instrumento o técnica de evaluación psicológica se le denomina erróneamente tests psicológico. Pero no todas las técnicas o instrumentos son tests puesto que para serlo precisan estar estandarizados y tipificados. Por el contrario a los tests sí les podemos denominar instrumentos de evaluación psicológica.

Según **Cohen y Swerdlik** la **prueba** es un dispositivo o procedimiento de medición diseñado para medir variables relacionadas con la psicología Ej: inteligencia, personalidad... Según **Gregory**, una prueba es un procedimiento estandarizado para tomar una muestra de conducta y describirla con categorías o puntuaciones. Gregory utiliza el término prueba como sinónimo de test y señala que las pruebas son sumamente variadas en sus formatos y aplicaciones, contemplando la mayor parte de éstas las siguientes características: procedimiento estandarizado, muestra de conducta, puntuaciones o categorías, normas o estándares y predicción de la conducta fuera de la prueba.

Cronbach define test como "procedimiento sistemático para observar el comportamiento y describirlo con la ayuda de escalas numéricas o categorías fijas". Por sistemático quiere decir que el examinador recoge la información interrogando u observando a todas las personas de la misma manera y en una situación idéntica o similar. Y añade que un test se considera estandarizado cuando las instrucciones del examinador, los aparatos y las reglas de corrección han sido fijadas de manera que las puntuaciones registradas en diferentes ocasiones son completamente comparables.

Solamente puede considerarse test a **aquellos instrumentos que están estandarizados y tipificados,**

y por lo tanto, nos informan de la puntuación de un sujeto en relación a otro o a su grupo de referencia(Fig. 3.1.).

2. CLASIFICACIÓN DE LOS INSTRUMENTOS DE EVALUACIÓN PSICOLÓGICA

Pervin organizó los instrumentos de evaluación en test proyectivos, test subjetivos, test psicométricos y test objetivos. En esta clasificación se utiliza la palabra test para las cuatro categorías sin tener en cuenta si están o no estandarizados y tipificados. **Fernández Ballesteros** realiza una clasificación en 6 categorías: técnicas de observación, técnicas objetivas, técnicas de autoinforme, la entrevista, técnicas subjetivas y técnicas proyectivas. Esta autora utiliza el término "técnica" para realizar una clasificación de los instrumentos de evaluación, y diferencia entre tests y técnicas de evaluación, e indica que "**la técnica de evaluación tan sólo supone dispositivos de recogida de información, sin que necesariamente requiera tipificación de su material o con ella se permitan comparaciones intersujetos**". La entrevista la sitúa aparte debido a que se trata del más importante y extendido de los autoinformes.

Esta misma clasificación de las técnicas de evaluación psicológica la comparten Forns, Abad, Amador, Kirchner y Roig y la relacionan con los distintos modelos de evaluación psicológica:

- Desde una **perspectiva biologicista**, las técnicas de estudio de la conducta humana son las técnicas objetivas de tipo psicofisiológico.
- Desde una **perspectiva psiquiátrica**, el estudio de síntomas se realizará con la entrevista dirigida, complementada con el uso de análisis psicofisiológicos, si fuera necesario.
- Las **técnicas proyectivas y la entrevista libre** son las prioritarias desde posturas psicoanalíticas y psicodinámicas.
- El **modelo fenomenológico** resalta la importancia de la propia experiencia y vivencias personales, serán utilizadas las técnicas subjetivas y de entrevista no directiva.
- El **modelo behaviorista** enfatiza el análisis de las conductas objetivables y el uso de la cuantificación, las técnicas apropiadas son la entrevista, la observación, las técnicas objetivas y los autoinformes.

En la actualidad la mayoría de psicólogos tienden a utilizar en el proceso de evaluación psicológica instrumentos desarrollados desde otros enfoques teóricos. Las técnicas de evaluación psicológica también pueden clasificarse en función de su **aplicación a lo largo del proceso de evaluación psicológica**. A medida que avanza el proceso se seleccionan distintos tipos de instrumentos.

Fernández Ballesteros define el **autoinforme** como "mensaje verbal que un sujeto emite sobre cualquier tipo de manifestación propia". Incluye los tests psicométricos entre los autoinformes, ya que suponen un informe verbal sobre la conducta y se consideran tipificados por estar construidos a través de procedimientos psicométricos, y agrupa a los autoinformes en base a la clase de variable que miden:

- **Rasgos, dimensiones o factores de personalidad**, como el MMPI
- **Estados como el STAI** (cuestionario de ansiedad rasgo/estado)
- **Repertorios clínicos conductuales** que informan sobre la conducta motora, cognitiva y fisiológica consideradas como muestras y no como rasgos intrapsíquicos
- **Repertorios, procesos y estructuras cognitivas**, que se trata de autoinformes sobre creencias, atribuciones, automensajes o autoinstrucciones que se da al sujeto frente a la situación problema en la que se encuentra o también autoinformes sobre el funcionamiento motivacional del sujeto.

Los **principales tipos de autoinformes** según Fernández Ballesteros son: la entrevista, los cuestionarios, inventarios y escalas, los autorregistros y los pensamientos en voz alta. Consideramos que a excepción de las pruebas subjetivas y las proyectivas, que utilizan materiales enmascarados, el resto de técnicas pueden llegar a ser un tipo de autoinforme.

Por otro lado, las **variables que generalmente miden los cuestionarios, inventarios y escalas** son la personalidad, los repertorios clínico-conductuales y de constructos cognitivos y motivacionales. Las técnicas de "pensamiento en voz alta" se aplican en la evaluación de conductas generalmente cognitivas. Así, los autoinformes pueden medir diversos tipos de variables. Debemos aclarar también la distinción entre cuestionario, inventario y escala. El **cuestionario** incluye una lista de cuestiones o preguntas, por lo tanto la formulación de los ítems se hace siempre con interrogación. Los **inventarios** se construyen a partir de un listado de ítems en forma de conjunto de frases que representan situaciones, conductas o respuestas a las que el sujeto tiene que contestar con qué frecuencia le ocurren. Y la **escala** comporta la observación externa.

Las **características básicas de clasificación de las técnicas de evaluación psicológica** son, según

Gregory:

- **Procedimientos estandarizados y no estandarizados**. Una prueba está estandarizada cuando tiene instrucciones fijas para su aplicación y calificación y se aplica a un grupo representativo de la población, para quienes está especialmente dirigido. En ellas se proporcionan normas o estándares. Las puntuaciones obtenidas se interpretan comparándolas con la muestra de estandarización. Los test estandarizados son pruebas referidas a la norma. Los procedimientos no estandarizados no poseen normas y por tanto no necesita comparar al individuo particular con un grupo de referencia; su objetivo es determinar la posición del sujeto evaluado con respecto a los objetivos. Los no estandarizados son pruebas referidas al criterio.

- **Pruebas individuales o grupales.** Un instrumento individual sólo se aplica a un sujeto, mientras que las pruebas grupales pueden aplicarse simultáneamente a varios sujetos.

- **Pruebas referidas a la norma o a criterio.** En la prueba referida a la norma, la puntuación de cada sujeto se interpreta con referencia a una muestra de estandarización, mientras que las pruebas referidas al criterio no necesitan de la comparación con el grupo de referencia, sino determinar la posición de cada sujeto con respecto a un criterio. El centro de atención se coloca en aquello que el examinado puede hacer. Las pruebas referidas al criterio identifican el dominio o falta del mismo del sujeto en relación con conductas específicas.

Forns y colbs presentan la siguiente clasificación de las técnicas:

- **El grado de estructuración de los estímulos y la respuesta.** Una prueba estructurada en el estímulo tiene una única interpretación, mientras que una prueba con menor grado de estructuración ofrece más variedad de interpretaciones.

- **El grado de enmascaramiento del objetivo de la prueba.** Una prueba no enmascarada deja claro desde el principio los objetivos que persigue, mientras que los instrumentos enmascarados poseen un objetivo distinto del que pretenden aparentar en un principio. En este grupo podemos situar a las técnicas subjetivas y las proyectivas.

- **El grado de inferencia interpretativa.** Los niveles de inferencia son propuestos por Sundberg, Tyler y Taplin y se refieren a las respuestas que un sujeto emite frente a cualquier dispositivo de evaluación o ante cualquier respuesta del sujeto. Son cuatro los niveles de inferencia que proponen estos autores y se ordenan de menor a mayor grado de abstracción: 1) **nivel I**, la conducta del sujeto es entendida como muestra de su comportamiento en la vida real. Supone un nivel mínimo de inferencia, 2) **nivel II**, la conducta evaluada del sujeto se asocia con otras conductas no evaluadas. Se apoya por tanto en un supuesto de relación, 3) **nivel III**, la conducta del sujeto expresa la existencia de un atributo subyacente en el sujeto, de carácter intrapsíquicos e inobservable, y 4) **nivel IV**, la conducta evaluada es una explicación especulativa a partir de una teoría concreta del psiquismo, y el concepto inferido se integra en una teoría completa. Las pruebas que aceptan un mayor grado de inferencia son las proyectivas y las cognitivas, y las que aceptan un menor grado de inferencia son las conductuales radicales y las conductual-cognitivas.

- **El grado de modificabilidad de la respuesta.** Se refiere al grado en el que el sujeto puede modificar o alterar su respuesta en la prueba. Así, las pruebas objetivas son las menos susceptibles de ser alteradas, mientras que en las técnicas de autoinforme el sujeto puede falsear con más facilidad su respuesta.

Además de estas agrupaciones, los instrumentos de evaluación también pueden clasificarse según **Aiken**a partir de sus contenidos verbal/no verbal, test de ejecución, o una prueba cognoscitiva o afectiva (Fig. 3.2.).

3. ¿QUIÉNES PUEDEN APLICAR UN INSTRUMENTO DE EVALUACIÓN PSICOLÓGICA?

En 1950 un Comité sobre Normas Éticas para la Psicología de la APA publicó un artículo en el que se definía tres niveles de pruebas en función del grado de conocimientos que su uso requería. Existe un proyecto de la Asociación Europea de Evaluación Psicológica EAPA para restringir el acceso de tests no sólo a los no psicólogos, sino también a los psicólogos no formados en evaluación o aquellos que no se reciclen periódicamente. Sin embargo, esto no siempre es así.

En el proceso de evaluación psicológica, cuando el objetivo de la demanda es la intervención psicológica, se administran pruebas en distintas fases del proceso, y se aplica un tratamiento psicológico, utilizando repetidamente y en distintos momentos las mismas pruebas para contrastar los beneficios del tratamiento y observar objetivamente los logros conseguidos. Tanto para la construcción de los instrumentos de evaluación como en su uso responsable, se

han elaborado Guías. Muñiz describe los aspectos éticos y deontológicos de la evaluación psicológica, explicando las normas generales que deben regir en la construcción de tests, en la práctica clínica, en la investigación psicológica, así como lo que debe saber un psicólogo para utilizar los tests adecuadamente, y señala algunos problemas actuales. Este autor resalta la clasificación de los instrumentos de evaluación en tres niveles (a, b, c), siguiendo las normas de la APA, asumidas por el Colegio Oficial de Psicólogos español:

- A) Formación y experiencia en el ámbito concreto de aplicación. Incluye instrumentos que pueden aplicarse, corregirse e interpretarse con sólo la ayuda del manual, por ejemplo, los test de rendimiento.
- B) Conocimiento sobre la teoría de los tests y métodos estadísticos, por lo que requieren formación técnica sobre construcción y uso de tests, así como de aspectos psicológicos, estadísticos, sobre diferencias individuales, personalidad, etc. Ej.: cuestionarios de personalidad.
- C) Titulación en psicología, psiquiatría o psicopedagogía y experiencia profesional en diagnóstico clínico, que requieren una preparación profunda de los tests y técnicas subyacentes, así como experiencia en su aplicación, por ejemplo, las técnicas proyectivas y las escalas de aplicación individual.

4. ¿CÓMO ELEGIR EL MEJOR INSTRUMENTO DE EVALUACIÓN PSICOLÓGICA?

La elección de las herramientas psicológicas más adecuadas, depende de **qué** se quiere evaluar, **para qué**y **quién o quiénes** van a ser evaluados. Debemos elegir aquellas pruebas que respondan a las necesidades específicas de la evaluación que vayamos a realizar y que posean la mayor fiabilidad y validez. Sin embargo, en muchas ocasiones, dispondremos de más de una prueba con buenas calidades psicométricas que evalúan un mismo constructo, y debemos elegir entre una de ellas. ¿Cómo la elegiremos? Lo primero que deberemos hacer es:

1. **Saber cuáles son las pruebas de inteligencia estandarizadas**, que son aquellas que tienen instrucciones específicas para su aplicación y calificación.
2. **Elegir aquellas que posean unos adecuados criterios de calidad:** fiabilidad y validez.
3. **Seleccionar aquellos instrumentos que puedan aplicarse a la edad del sujeto** o sujetos que queremos evaluar.
4. **Seleccionar un instrumento de administración individual o grupal**, según sea nuestro caso, teniendo en cuenta además otros aspectos como el tiempo de aplicación, formato de prueba y el objetivo que se pretende evaluar.

Una vez elegida la prueba/s, es imprescindible que el evaluador:

1. Se familiarice con la prueba.
2. Prepare adecuadamente el lugar donde se aplicará.
3. Cree un ambiente y rapport adecuados.
4. Prepare los materiales necesarios.
5. Explique el propósito de la evaluación antes de aplicar la prueba y cómo se va a utilizar la información obtenida.
6. Siga estrictamente las normas de aplicación.
7. Corrija las pruebas siguiendo los pasos que se indican en el manual.
8. Cumpla con las obligaciones éticas y deontológicas antes de iniciar la evaluación, solicitando el consentimiento del propio sujeto o su representante legal y clarificando quiénes van a tener acceso a la información obtenida.

En síntesis, la mejor prueba se elegirá teniendo en cuenta los criterios mencionados, así como será imprescindible la destreza del evaluador, por lo que el psicólogo deberá adquirir previamente experiencia en su administración y no aplicar un instrumento hasta que posea una plena seguridad y conocimiento del mismo (Fig. 3.3.).

5. ¿CÓMO SON LOS MANUALES DE APLICACIÓN DE LOS INSTRUMENTOS DE EVALUACIÓN PSICOLÓGICA?

Recordemos que una prueba está estandarizada cuando tiene unos procedimientos claramente definidos para su administración y corrección. En los manuales se incluyen instrucciones para su aplicación y los datos del

grupo normativo con el fin de comparar la puntuación obtenida por el sujeto evaluado con el grupo de referencia. Las partes de que consta un manual son: introducción, objetivos, descripción general que incluye una ficha técnica con la descripción de la prueba, fundamentación estadística, instrucciones para la aplicación, corrección e interpretación, ventajas y limitaciones de la prueba y áreas de aplicación e investigaciones recientes realizadas con ese tests.

El buen evaluador debe leer a fondo el manual antes de aplicar el test y prestar una atención especial a los siguientes aspectos:

- Sobre la **construcción de la prueba**, debe estar atento a lo que mide, para qué sirve, a qué tipo de población va dirigida, si describe la muestra normativa, indica el proceso de creación de la prueba, sus revisiones, si las ha habido...
- El manual describe detalladamente **cómo debe realizarse la administración del test**, las instrucciones que deben darse para su ejecución, el tiempo de aplicación máximo, la puntuación que se otorga a las respuestas del sujeto...
- El manual describe con claridad los **pasos a seguir para corregir y obtener los resultados de la prueba**.

5.1. ¿Cómo realizar una correcta administración de las pruebas de evaluación psicológica?

Al sujeto se le deben decir **las palabras exactas indicadas en el manual** y no una interpretación de las mismas. Cuando el evaluado solicite una aclaración, podrá dársela o no si lo permite el manual. Por lo general el autor tiene previstos algunos de los casos que con mayor frecuencia se pueden encontrar el evaluador, indicando en el manual la forma adecuada de proceder.

Cuando el examinador tenga poca práctica en la aplicación es recomendable que lea textualmente las instrucciones que deben darse al sujeto para la ejecución de cada prueba. Es importante también que esté atento a los tiempos máximos de ejecución de las pruebas. Este aspecto, junto a la observación de cómo realiza la tarea, aporta información cualitativa a la que debe estar atento el evaluador y que le será de gran utilidad si el objetivo es diseñar un programa de intervención psicológica.

Para la corrección y puntuación de las pruebas psicológicas, se deberán seguir las indicaciones correspondientes. Primero se realizarán las acciones oportunas para conocer la puntuación directa del test. Esta puntuación no nos informa todavía de los resultados que ha obtenido el sujeto, y puede ser malinterpretada debido a que no puede considerarse hasta que no se consultan los baremos del test y se transforma la puntuación directa obtenida en otro tipo de valores que son los que van a permitir comparar al sujeto evaluados con su grupo de referencia en la característica evaluada.

Cuando la prueba permite su corrección a través del ordenador es recomendable su utilización. Otra ventaja que ofrece la corrección automatizada es la economía de tiempo de los evaluadores y la capacidad de analizar grandes cantidades de datos y compararlos de forma simultánea con otros en su memoria.

6. CRITERIOS DE CALIDAD EXIGIBLES A LOS INSTRUMENTOS DE EVALUACIÓN PSICOLÓGICA

Los principales criterios psicométricos de calidad o bondad asumidos como normas en la construcción, interpretación y utilización de instrumentos psicológicos de medición son la **fiabilidad** y la **validez**.

6.1. Fiabilidad

La APA describió la **fiabilidad** como la exactitud de la medición de un test, es decir, la precisión con la que mide la prueba. La definición de los Standards for Educational and Psyuchologicla Testintg resalta que la fiabilidad se refiere al grado en que los resultados del examen son atribuibles a fuentes sistemáticas de varianza. Una década más tarde se designa la fiabilidad como el grado en que las puntuaciones del test son consistentes, dependientes, o repetibles, es decir, el grado en que están libres de errores de medida. El cálculo de la fiabilidad nos informa de la cuantía de error de un instrumento de medida, por lo que, a menor error, mayor fiabilidad, y más exacto o preciso será el test.

Desde la teoría clásica de los tests, **Aiken** explica que se supone que la puntuación observada que obtiene una persona en una prueba se compone de una clasificación real más algún error no sistemático de medida. La**calificación real** se define como el promedio de las calificaciones que se obtendrían si una persona realizara la prueba una cantidad infinita de veces. Enfatiza que la calificación real nunca puede medirse con exactitud, sino que debe calcularse a partir de la calificación observada que obtuvo la persona en la prueba. También se supone que la varianza de las calificaciones observadas para un grupo de sujetos es igual a la varianza de sus calificaciones reales más la varianza de errores no sistemáticos de medición. Así, la fiabilidad de la prueba se define como **la relación de la varianza real con la varianza observada** o la proporción de la varianza observada que se explica por la varianza real.

El coeficiente de fiabilidad es un índice de confianza, por lo que no es un valor de todo o nada, sino que existen distintos tipos y grado de fiabilidad. Se supone que una puntuación en una prueba de capacidad refleja tanto la puntuación verdadera de quien responde la prueba en la capacidad que se está midiendo como el error. La falta de fiabilidad es el resultado de los errores en la medida que se producen por estados internos temporales, como baja motivación o indisposición, o condiciones externas, como un entorno incómodo o con distractores para una prueba.

Gregory resalta que muy pocas medidas de las características físicas o psicológicas son totalmente consistentes, incluso de un momento al siguiente. Según este autor es mejor considerar el concepto de fiabilidad como un continuo que abarca desde la consistencia mínima de una medición a la casi perfecta repetibilidad de los resultados.

Así, debemos exigir una alta fiabilidad en los instrumentos de evaluación que seleccionemos. Así, los niveles de fiabilidad alta (superiores a 90) son necesarios cuando se han de tomar decisiones que afecten a individuos. Los test de fiabilidad moderada (75-85) pueden ser utilizados como pruebas preliminares o de cribado. Las pruebas de fiabilidad baja (inferior a 65) han de ser rechazadas, ya que incluyen un exceso de error.

6.2. Fuentes de varianza de error

Las **principales fuentes de varianza de error** son:

- **Construcción de pruebas**. En la construcción de una prueba se puede generar una fuente de varianza en el **muestreo de reactivos o muestreo de contenidos**. Si se comparan dos o más pruebas que midan una misma capacidad, atributo... se verá que el número de elementos es distinto, además están redactados de forma diferente. Un desafío en la elaboración de una prueba es maximizar la proporción de varianza total que es invarianza verdadera y minimizar la proporción de la varianza total que es varianza de error. En una prueba bien diseñada, el error de medición proveniente de la muestra de reactivos será mínimo y una prueba siempre constituye una muestra y nunca la totalidad del conocimiento o conducta de una persona.
- **Administración de pruebas**. Durante la aplicación de la prueba pueden desencadenarse fuentes de varianza de error que pueden incluir y modificar la atención y motivación del sujeto evaluado. Algunas de estas fuentes pueden estar relacionadas con el ambiente de aplicación, otras son las relativas al sujeto evaluado. El evaluador también puede contribuir a las fuentes de variación, con una incorrecta apariencia física, un comportamiento y profesionalidad inadecuados... Por lo tanto un test puede ser fiable desde el punto de vista psicométrico, pero fallar por elementos ajenos a él.
- **Calificación e interpretación de las pruebas**. La corrección de las pruebas por ordenador o mediante lectura óptica elimina la varianza de error al no cometer fallos en la calificación, y por consiguiente, en su interpretación. Sin embargo, todavía son muchas las pruebas que el psicólogo debe corregir manualmente, pudiendo convertirse en una fuente de varianza de error cuando no se realiza correctamente.

Así, las pruebas deben disponer de criterios de corrección lo más objetivos posible.

6.3. Tipos de Fiabilidad

6.3.1. Coeficiente test–retest o estabilidad del test

Se halla al correlacionar las puntuaciones que obtiene un grupo de sujetos en la aplicación de una prueba con las obtenidas en una segunda aplicación. Se espera que los sujetos obtengan puntuaciones semejantes en el mismo test aplicado en dos momentos distintos. Este tipo de fiabilidad tiene en cuenta los errores de medida derivados de las posibles diferencias de las condiciones en las que en dos ocasiones se ha aplicado la misma prueba. Pero, no refleja los errores relativos a distintas muestras de reactivos o elementos de la prueba. Si el intervalo de tiempo entre test-retest es pequeño, la fiabilidad será mayor que si se aumenta el tiempo entre ambos pases. Suele recomendarse unos seis meses como máximo entre el primer y el segundo estudio.

6.3.2. Coeficiente de formas paralelas/alternas o de equivalencia

En el cálculo del coeficiente test-retest la fiabilidad aumenta cuanto menor es el tiempo que ha transcurrido entre ambos, sin embargo, esto afecta a las respuestas en el segundo pase de la prueba, pues los sujetos recordarán el contenido de la misma. Esto no sería un problema si lo recordaran de la misma forma, pero las diferencias individuales harán que unos recuerden unos elementos más que otros, reduciendo la correlación entre ambas aplicaciones.

El coeficiente de formas paralelas o de equivalencia consiste en **aplicar la segunda vez una forma paralela o alterna del test** y de esta forma se evitarán dos tipos de errores: 1) los debidos a distintos reactivos y 2) los errores derivados de las dos ocasiones diferentes de aplicación. Pero no todos los autores comparten esta opinión. Gregory indica que el coeficiente de formas alternas **introduce diferencias en la muestra de reactivos**,

debido a que algunas personas pueden tener un mejor o peor desempeño en una forma de la prueba, dado la muestra particular de reactivos, lo que no ocurre en el coeficiente test-retest porque se utilizan los mismos reactivos en ambas ocasiones.

Aiken describe el procedimiento correcto. Se trata de elaborar dos formas de la misma prueba y aplicar en el primer pase de la prueba la forma A a la mitad del grupo y la forma B a la otra mitad. Y en la segunda aplicación invertirlo. La correlación que resulte entre las calificaciones de las dos formas se conoce como **coeficiente de estabilidad y equivalencia**, y según Aiken, tiene en cuenta tanto los errores debidos a distintos momentos de aplicación, como los distintos reactivos de la prueba.

6.3.3. Coeficiente de consistencia interna

El coeficiente de consistencia interna es más sencillo y **tiene en cuenta los errores de diferentes muestras de reactivos de una prueba, pero no refleja los errores de medición debidos a las diferentes condiciones o momentos de aplicación**. Puede calcularse a través de distintos métodos estadísticos: método de división por mitades, método de Kunder-Richardson y coeficiente alfa de Cronbach. El coeficiente alfa es el método estadístico preferido para obtener una estimación de la fiabilidad y de la consistencia interna en una prueba.

6.3.4. Coeficiente interjueces o entre evaluadores

La fiabilidad entre evaluadores es el grado de acuerdo o consistencia que existe entre dos o más evaluadores. Según Aiken para determinar la fiabilidad interjueces dos personas califican las respuestas de un grupo de sujetos y después se calcula la correlación entre los dos grupos evaluados. Otro planteamiento es hacer que varias personas califiquen las respuestas de un sujeto a la prueba, o hacer que varias personas califiquen las respuestas de varios sujetos. Este último planteamiento produce un **coeficiente entre clases o coeficiente de concordancia** que es un coeficiente de fiabilidad entre calificadores generalizado. El cálculo de fiabilidad entre intercalificadores es sencillo. Dos o más examinadores califican de manera independiente una muestra de las pruebas y entonces se correlacionan las puntuaciones por pares de examinadores. Este tipo de fiabilidad complementa otras estimadas, pero no las sustituye.

6.4. Validez

La definición de **validez** indica que una prueba es válida al grado en que las inferencias que se realicen a partir de ella sean apropiadas, significativas y útiles. Según Cronbach lo que se evalúa no es el instrumento, sino la interpretación de los datos que se obtienen del mismo. La validez no es una propiedad del test o de la evaluación como tal, sino más bien el significado de las puntuaciones. Según Aiken, una prueba puede caracterizarse por muchos tipos de validez, dependiendo de los propósitos específicos con los que se diseñó, la población a la que se dirige y el método para determinar dicha validez. Como hemos visto, la fiabilidad puede estar influida por errores de medida no sistemáticos. La validez de una prueba se puede ver afectada tanto por errores no sistemáticos como por errores sistemáticos que hacen referencia a que, a pesar de que una prueba se desarrolla con la finalidad de evaluar un constructo determinado, es muy difícil valorar un rasgo aislado sin la influencia de otros, por lo que el error sistemático de medición surge cuando la prueba mide de manera consistente alguna otra variable que no es el rasgo para el cual se creó. Por ello una prueba puede ser fiable sin ser válida, pero no puede ser válida sin ser fiable. Silva hace algunas **matizaciones sobre la validez**:

- La validez está relacionada con las inferencias que se hagan a partir de las puntuaciones obtenidas mediante un instrumento en determinadas circunstancias.
- No se valida el instrumento, sino las interpretaciones que se hagan a partir de sus puntuaciones.
- La validez es algo estimado, algo que se infiere a partir de un conjunto de informaciones y no algo que se reduce a un coeficiente o coeficientes particulares.
- No debe hablarse de tipos o clases de validez, sino de tipos o clases de evidencia. El concepto de validez es esencialmente unitario.

Aiken y Cohen y Swerdlik indican que los **métodos mediante los cuales pueden evaluarse la validez**son:

- El análisis del contenido.
- La relación de las puntuaciones obtenidas en la prueba con las puntuaciones en base a un criterio de interés u otras medidas.
- El análisis general de las características psicológicas o constructos particulares que mide la prueba.

Estos tres enfoques no son mutuamente excluyentes para la evaluación de la validez, cada uno contribuye a un juicio de la validez de prueba y proporciona un panorama unificado de la validez de la prueba.

6.4.1. Validez de contenido

Representa la **comprobación de que el contenido de la técnica en cuestión comprenda una muestra representativa del universo posible de conductas que se pretende evaluar con ella**. Se relaciona con el enfoque referido a criterios y considera a un test como una muestra de un conjunto definido de conductas. Una definición que clarifica el propósito de validez de contenido es la que ofrece **Lennon**: la validez de contenido se refiere a las respuestas del sujeto más que a las preguntas mismas del test, con el fin de enfatizar el hecho de que la estimación de la validez de contenido debe tomar en cuenta no sólo el contenido de las respuestas, sino también el proceso que presumiblemente emplea el sujeto para llegar a su respuesta.

El análisis de validez de contenido se aplica más frecuentemente en **pruebas de conocimiento o rendimiento**, y se compara con el contenido de la prueba con los objetivos de los conocimientos o rendimientos escolares del nivel escolar que se está midiendo. La validez de contenido mejora cuando se planifica el test cuidadosamente, y requiere una visión clara de lo que éste pretende medir y debe cubrir los siguientes aspectos: un rango apropiado de tareas, estímulos y/o situaciones, la clase de respuesta que el observador registra y las instrucciones que informan al examinado de lo que tiene que hacer. También se tiene en cuenta en las medidas de aptitud, interés y personalidad.

6.4.2. Validez criterial

La validez criterial, también llamada predictiva, **expresa el grado en que las puntuaciones en una variable, usualmente un predictor, pueden utilizarse para inferir el rendimiento en una variable diferente y operacionalmente independiente llamada criterio**. La variable que debe ser predicha es la criterio, p. ej, el rendimiento académico, y el predictor, aquella a través de la cual se predice, p. ej, un test de inteligencia, y la validez criterial expresaría la "convergencia de indicadores". Dos tipos de evidencia se incluyen bajo la denominación "validez con base a criterios".

Una es la **validez concurrente**, que es la forma de validez relacionada con un criterio que es un índice del grado en que una puntuación de una prueba se relaciona con alguna medida criterio obtenida al mismo tiempo. Ej: el diagnóstico psiquiátrico actual de los pacientes sería una medida apropiada de criterio para proporcionar evidencia de validez para una prueba psicodiagnóstica de papel y lápiz. Es frecuente que las correlaciones entre una nueva prueba y otras existentes se citen como evidencia de validez concurrente. Para realizar este tipo de validez, las pruebas antiguas deben satisfacer dos condiciones: la primera es que las pruebas criterio deben haberse validado a través de correlaciones con datos conductuales apropiados que no se hayan obtenido con pruebas. En segundo lugar, el instrumento a validar debe medir el mismo constructo que las pruebas criterio.

La otra es **la validez predictiva**, que es la forma de validez relacionada con un criterio que es un índice del grado en que una puntuación de una prueba predice alguna medida criterio. En este tipo de validez las medidas de criterio se obtienen en el futuro. Ej: las calificaciones universitarias pronosticadas a partir de un examen de ingreso.

Existen una serie de **factores que pueden afectar a la validez criterial**:

- **Diferencias de grupo**: las variables moderadoras de edad, sexo y rasgos de personalidad pueden afectar la correlación entre una prueba y una medida de criterio. Los coeficientes de validez tienden a ser más reducidos en grupos más homogéneos. Una prueba que representa un indicador válido de una variable criterio particular en un grupo de sujetos debe tener validez cruzada, que comprende la aplicación de la prueba a una segunda muestra de personas para determinar si conserva su validez en distintas muestras.

- **Extensión de la prueba**: al igual que la fiabilidad, la validez varía en función de la extensión de una prueba y la heterogeneidad del grupo de personas que la presenta. Las puntuaciones obtenidas en pruebas extensas y que se apliquen a un grupo de sujetos que varíen en gran medida en las características a medir tendrán varianzas mayores.

- **Contaminación de criterios**: a veces el criterio se distorsiona debido al método particular para determinar las calificaciones de criterio. El método de comparar grupos, provocará evidencias falsas para la validez de la prueba. Esta contaminación puede controlarse a través del análisis a ciegas, es decir, sin comunicar a quien realiza el diagnóstico ninguna información sobre los sujetos parte de las calificaciones de la prueba. Pero no todos los psicólogos están de acuerdo.

- **Índice de base**: se refiere a la proporción de personas en la población que muestran la característica o comportamiento de interés.

- **Incremento de la validez**: éste se refiere a que aumenta la precisión de las predicciones y los diagnósticos cuando el instrumento se incluye en una batería de

técnicas de evaluación, frente a las ocasiones en que se utiliza separadamente.

6.4.3. Validez de constructo

La validez del constructo establece el **grado en el cual un instrumento mide o guarda relación con un determinado rasgo o constructo hipotético**. Algunos autores afirman que toda medición debería referirse a constructos, debido a que integra las consideraciones criteriales y de contenido. Muchos autores consideran la validez de constructo como unificador de los tipos de evidencia de validez. Silva propone 10 características más importantes de la validez de constructo, algunas de las cuales agrupamos para poder diferenciar los conceptos de constructo y validez de constructo.

- **Constructo**
    - es sinónimo de concepto científico
    - no debe ser considerado como algo estático
    - tanto los constructos como la validación de constructo, están indisolublemente ligados a la evidencia empírica, pero un constructo no se reduce a sus referentes empíricos, conserva siempre un excedente de significación.
    - Posee un estatus fundamentalmente epistemológico, es un medio de conocimiento.
    - No se propone sólo con fines especulativos, sino con el fin de potenciar la predicción.
    - Su valor se juzga por su utilidad.
- **Validez de constructo**
    - es sinónimo de validez conceptual o grado de adecuación de las inferencias conceptuales teóricas que se hacen a partir de los datos de evaluación.
    - Se refiere tanto al concepto como al método implicado.
    - Engloba en si los conceptos de validez criterial y de validez de contenido.
    - No existe límite en cuanto a las estrategias, procedimientos, instrumentos y tipos de datos potencialmente útiles.
    - No se expresa sólo en función de uno o algunos coeficientes, sino que se estima en función de toda la información acumulada en torno a las hipótesis planteadas.
    - Consiste esencialmente en la aplicación del proceso de formulación y contrastación de hipótesis científicas al campo de la evaluación psicológica.

La validez de constructo es un tipo de validez más general, no se determina de una sola forma o por medio de una investigación, sino que comprende un conjunto de investigaciones y procedimientos diseñados para determinar si un instrumento de evaluación que mide cierta variable cumple su cometido.

6.5. Relación entre fiabilidad y validez: un continuo de generalizabilidad

Los criterios psicométricos tradicionales de fiabilidad y validez no son aceptados por todos los autores conductuales, algunos piensan que son algo limitados. Con el fin de ofrecer una alternativa surge la **Teoría de la Generalizabilidad**, que supone una reconceptualización más amplia de los conceptos de fiabilidad y validez, en la que aparece el concepto de “puntuación universo” que expresa el grado de inferencia que el examinador realiza desde una muestra de datos observados a un conjunto de datos de interés procedentes de diferentes ámbitos. Así, los datos de un test tendrán interés por cuanto son muestras representativas del universo de datos que podrían ser obtenidos. Pero ¿hasta qué punto una observación puede generalizarse a otras observaciones?

Silva señala que la Teoría de la Generalizabilidad permite lanzar un puente conceptual entre finalidad y validez e indica que ambos se hallan sobre un continuo de generalizabilidad: la fiabilidad supone la relación de un test consigo mismo, por lo que se refiere a la generalizabilidad consigo mismo, mientras que la validez se relaciona con otra prueba, criterio o constructo, y por tanto la generalización va más allá del test.

6.6. Aplicaciones de la Teoría de Respuesta al Ítem (TRI)

La **TRI** ha reemplazado a la Teoría Clásica de medida como marco para el desarrollo de tests, construcción de escalas... Tanto en la teoría clásica de los tests como en la teoría de la generalizabilidad, las puntuaciones de un

test son más dependientes de la muestra que de la propia función analizada. La TRI trata de subsanar dos problemas. El primero hace referencia al **error en la medida** y asume que las puntuaciones de los sujetos en un test estarán afectadas por un error aleatorio, atribuible a diversas causas: dependientes del sujeto, del ambiente, del instrumento y del propio proceso de evaluación. El segundo se refiere a la **invarianza de las mediciones y las propiedades de los instrumentos**. Los **principales objetivos** de la TRI son:

- Búsqueda de medidas que sean independientes de las puntuaciones estándar derivadas del grupo.
- La elaboración de nuevas pruebas que analicen la invarianza de la conducta en sí misma, de modo que un test represente con precisión un dominio gradual de conocimiento relativo a una única medida.
- La relación de los dos conceptos anteriores permite un tipo de medida en la que los parámetros de ítem y de persona son ambos invariantes, de tal modo que ni la elección de una muestra de sujetos, ni la elección de los ítems afecte a los parámetros de dificultad del ítem ni a los de la habilidad.
- La agilidad en la combinatoria de ítems de test, que pertenezcan a un mismo dominio de conducta, dando paso a la aplicación de tests adaptados al sujeto, en función de la capacidad de las habilidades de cada individuo.

En cuanto al **cálculo estadístico**, la TRI utiliza un modelo matemático logístico para describir la relación entre el nivel de habilidad del examinado y la probabilidad que éste dé una respuesta correcta a un ítem del test. Algunas aplicaciones de la TRI han consistido en la creación de bancos de ítems y los diseños de tests a la medida del sujeto o test adaptativos computadorizados (TAC). Los test de medida consisten en la selección de informatizada de los ítems que puedan medir mejor la habilidad de un individuo.

7. PUNTUACIÓN DE LAS PRUEBAS DE EVALUACIÓN PSICOLÓGICA

7.1. Puntuaciones directas

Las puntuaciones directas son **el resultado directo e inmediato que se obtiene a la hora de corregir un test**. Gregory las denomina puntuación natural, ya que es el resultado inicial de la prueba y casi siempre resulta de la suma de los puntos otorgados a los aciertos del sujeto en un test. Estas puntuaciones no tienen significado por sí mismas, sino que lo adquieren cuando se comparan con algo, que puede ser un punto de referencia al criterio y/o a la norma.

7.2. Puntuaciones referidas al criterio

Una puntuación referida al criterio, o lo que es lo mismo, al universo de conductas, se interpreta en función de unos logros u objetivos a cumplir, arbitrariamente definidos, y que sirven para tomar decisiones. Se trata de una medida en términos absolutos que se refiere a un determinado grado de habilidad y a unos contenidos específicos.

Este tipo de puntuaciones nos informan acerca del **dominio que tiene un individuo en una habilidad particular**. Desde esta perspectiva se observan diferencias intraindividuales. Se centra en conocer aquello que el sujeto puede hacer y no en comparar con los niveles de ejecución de otros individuos y así identifican el dominio absoluto de la persona examinada atendiendo a conductas específicas.

Una de las principales aplicaciones de la evaluación referida al criterio es instruccional, se aplica generalmente en la evaluación educativa, y no necesita transformarse a otra puntuación debido a que tiene sentido en sí misma. Ej: cuando un sujeto ha acertado el 80% de las preguntas significa que ha adquirido el 80% de las competencias que se precisaban. Las principales características de las puntuaciones referidas al criterio son: a) los criterios de superación de la tarea son conocidos por el profesor y el estudiante y válidos para tomar decisiones, b) la ejecución del individuo se contrasta con la exigencia de la tarea, c) la ejecución provee información tanto de lo que el escolar domina como de lo que no, y d) la investigación provee la determinación de los puntos de corte en sujetos que dominan y los que no dominan la tarea.

Con referencia a un criterio, y en particular las pruebas de dominio, las diferencias individuales entre los examinados en las puntuaciones totales pueden ser mínimas. Sólo pueden servir en casos en los que pueden adoptarse estimaciones tradicionales.

7.3. Puntuaciones referidas a la norma

Una puntuación referida a la norma se interpreta a partir de un grupo de referencia, es decir, se basa en**la comparación de la ejecución de un sujeto con su grupo normativo**. La mayor parte de pruebas psicológicas

se interpretan a través de la consulta de normas. La puntuación que obtiene el sujeto indica la posición del mismo con respecto al grupo de referencia, y no tiene valor interpretativo propio sino que debe relacionarse con la norma que sustenta la medida. Para ello se transforma la puntuación que obtiene el sujeto en otra posición que indique la posición que ocupa respecto a ese grupo. Existen tres tipos fundamentales de puntuaciones normativas. La**puntuación percentil** indica el porcentaje de sujetos del grupo normativo que puntúan por debajo de la puntuación obtenida. La **puntuación cronológica** presenta la relación que guarda la puntuación en el test con la edad cronológica del sujeto. Y la **puntuación típica** señala la distancia que separa a un sujeto de la media del grupo normativo, expresando dicha distancia en unidades de desviación típica.

7.3.1. Puntuaciones percentiles

Sirven para **ordenar a los sujetos e indican el porcentaje del grupo que se deja por debajo**. Ej: un sujeto con un percentil 80 significa que obtiene puntuaciones superiores al 80% de los sujetos de su grupo de referencia, o que tiene puntuaciones inferiores al 20% restante. A pesar de que son fáciles de calcular, no permiten explicarlas diferencias entre percentiles ni permite comparara los percentiles obtenidos por un sujeto en distintos instrumentos de evaluación. Se trata de puntuaciones de orden, que en ningún caso ponen de manifiesto la diferencia cuantitativa que existe entre los individuos al no operar con unidades constantes. Son muy útiles en pruebas de rendimiento tanto a nivel educativo como empresarial.

7.3.2. Puntuaciones cronológicas

Permiten interpretar la puntuación que obtienen un sujeto en función de su edad. Así, se emplean en poblaciones infantiles y cuando se aplican tests de inteligencia general. Son básicamente dos. Por un lado, la**edad mental**, que es la puntuación media que obtienen en una prueba el conjunto de la población de esa edad. El problema es que un año de edad mental no significa lo mismo a lo largo del desarrollo. Por otra, el **cociente intelectual**. Elimina el problema anterior al dividir la edad mental por la edad cronológica y se define como la razón entre la edad mental y la edad cronológica multiplicada por 100.

7.3.3. Puntuaciones típicas

Las puntuaciones típicas nos indican **cuánto se separa el sujeto de la media del grupo de referencia, en función de lo que se separan los demás**. La puntuación directa que obtiene un sujeto se transforma en otra puntuación en relación a la media del grupo pero tomando como unidad de medida la desviación típica de ese grupo. El cálculo de las puntuaciones típicas puede presentar valores decimales y valores negativos, y para salvar estos inconvenientes, suelen realizarse **puntuaciones típicas derivadas**. Ej: la escala T.

Existen además las **puntuaciones típicas normalizadas** que han sido creadas mediante la normalización de la distribución original de las puntuaciones directas en el test.

7.4. Puntuaciones independientes de la norma

Estas puntuaciones se fundamentan en la Teoría de Respuesta al Ítem y facilitan la idea de unidimensionalidad de la habilidad analizada. Las puntaciones obtenidas en un test no precisan ser referidas a los resultados normativos de un grupo, sino que representan, en sí mismas, unos valores determinados en la dimensión de la aptitud analizada, reflejando adecuadamente el nivel de habilidad del sujeto. Tienen la ventaja de realizar un perfil individual y preciso de cada sujeto que muestre las áreas fuertes y débiles.