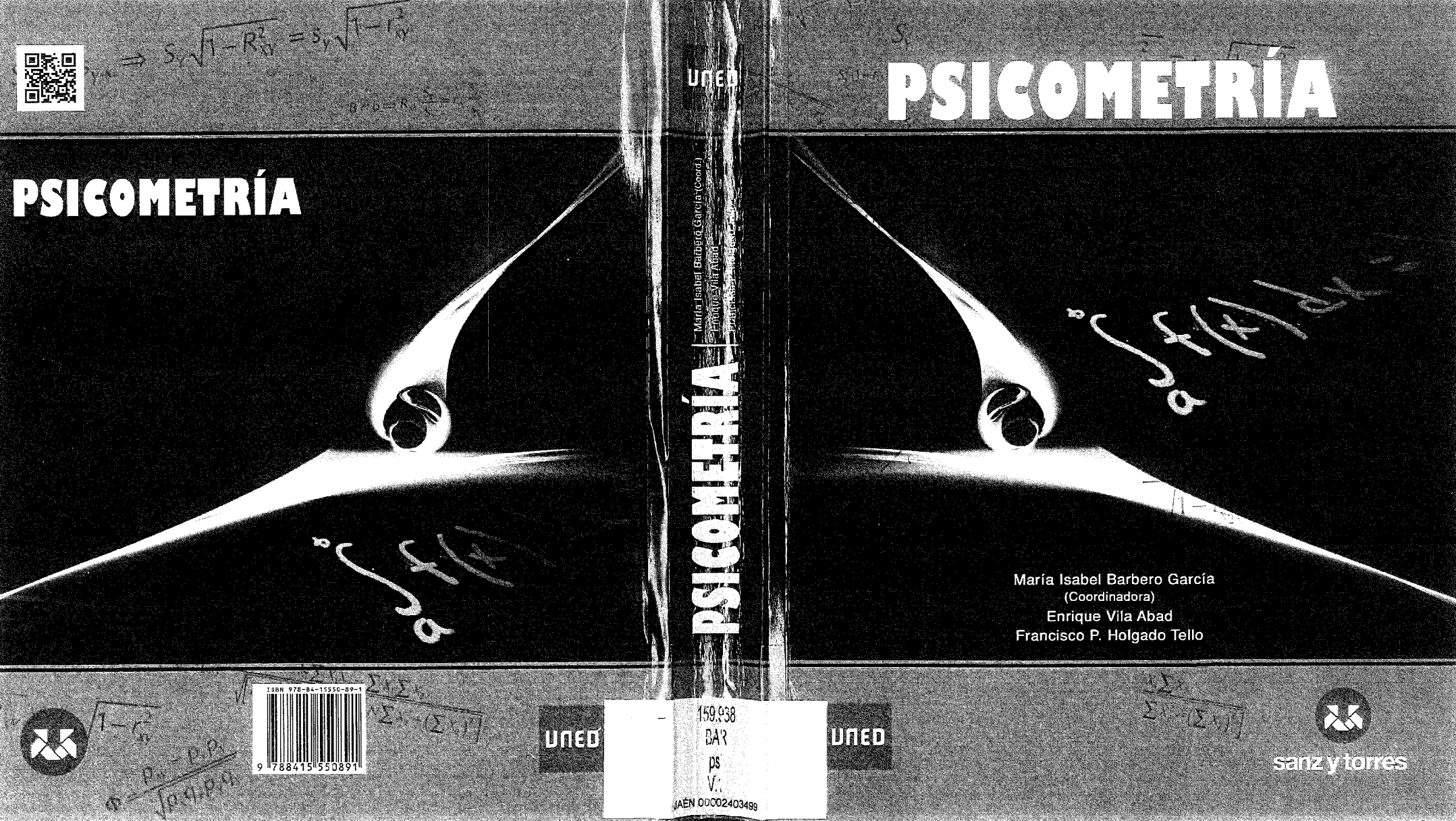


ISBN 978-84-15550-89-1

Colección

UNED

PSICOMETRÍA

© EDITORIAL SANZ Y TORRES, S. L.
Vereda de los Barros, 17
Pol. Ind. Ventorro del Cano – 28925 Alcorcón (Madrid)
☎ 902 400 416 – 91 3237110
www.sanzytorres.com
libreria@sanzytorres.com
www.sanzytorres.com/editorial
editorial@sanzytorres.com

ISBN (obra completa): 978-84-15550-89-1
ISBN: 978-84-15550-87-7
Depósito legal: M-21586-2015

Portada:
Javier Rojo Abuín

Composición:
Iván Pérez López

Impresión:
Medianil Gráfico, S. L., c/ Edison, 23, Pol. Ind. San Marcos, 28906 Getafe (Madrid)

Encuadernación:
Felipe Méndez, S. A., c/ Del Carbón, 6 y 8, Pol. Ind. San José de Valderas 2, 28918 Leganés (Madrid)

ÍNDICE

Parte I
CONSTRUCCIÓN DE INSTRUMENTOS DE MEDICIÓN PSICOLÓGICA

Parte II
EVALUACIÓN DE LAS PROPIEDADES MÉTRICAS DE LOS INSTRUMENTOS DE MEDICIÓN PSICOLÓGICA

Parte III
APLICACIÓN DE LOS INSTRUMENTOS Y EVALUACIÓN DE LOS SUJETOS

PRESENTACIÓN

La mayoría de las ideas fundamentales de la ciencia son esencialmente sencillas y, por regla general, pueden ser expresadas en un lenguaje comprensible para todos.

Albert Einstein

Cuando nos planteamos la realización de este libro pensamos que podría ser de utilidad no sólo a los alumnos que cursan el Grado de Psicología, como parte del material didáctico que deben utilizar para la preparación de la asignatura de Psicometría, sino a todas aquellas personas que, o bien por el trabajo que desarrollan, o simplemente por interés personal, están relacionadas con el tema que nos ocupa. No obstante, los que acapararon nuestro mayor interés y esfuerzo fueron los alumnos de Psicología de la Universidad Nacional de Educación a Distancia (U.N.E.D.), nuestros alumnos. Teniendo esto en cuenta, y considerando que el enfoque metodológico que subyace al Espacio Europeo de Educación Superior (EEES) y el modelo educativo del que parten los sistemas de Enseñanza a Distancia, ponen el acento en la actividad individual y en el trabajo en solitario de los alumnos, es fácil deducir que este material didáctico constituye uno de los elementos fundamentales en los que se apoyan este tipo de sistemas y, por lo tanto, es necesario atender, en su elaboración, tanto a su calidad científica como a su calidad pedagógica para, de esa manera, conseguir que nuestros alumnos encuentren más fácil y motivadora la tarea que deben emprender.

No se trata de un libro de texto en el sentido tradicional, ya que éstos están pensados para servir de apoyo a las explicaciones proporcionadas en clase por los profesores, se trata de un libro pensado para los alumnos de la UNED que se presenta como una alternativa a dichas explicaciones; por lo tanto, es el resultado de una adecuada y cuidadosa planificación, respondiendo en su organización, a los distintos momentos del acto de aprendizaje que tiene que realizar el alumno, e incluyendo las claves y fuentes de información complementarias para facilitar la adquisición de conocimientos.

Con el fin de que nuestros alumnos encuentren más fácil el estudio de los temas que se incluyen, se ha procurado utilizar un lenguaje sencillo y numerosos ejemplos que ayuden a la com-

prensión de los conceptos que en ellos se estudian. Además, al principio de cada tema se incluye un apartado con una serie de orientaciones didácticas en las que se exponen los puntos básicos que se van a ir analizando a continuación, y al final de cada tema se incluyen una serie de ejercicios de autoevaluación a través de los cuales los alumnos podrán comprobar el nivel de conocimientos adquirido, tanto a nivel teórico como práctico, así como una bibliografía complementaria para que pueda ser utilizada por aquellos que deseen completar su formación.

Cualquier experto que lea el libro podrá pensar que no se hace en él ninguna aportación novedosa ya que los contenidos que incluye se pueden encontrar en otros libros. Eso es cierto pero, tal y como hemos señalado anteriormente, se trata de manuales de texto que a veces les resultan difíciles de comprender a nuestros alumnos al enfrentarse en solitario a su lectura; por eso, creemos que es necesario ofrecerles éste especialmente pensado para ellos.

Ante la imposibilidad material de abordar, en un curso académico, la totalidad de los campos y conocimientos a que hace referencia la Psicometría, consideramos que era imprescindible llevar a cabo un proceso de selección de contenidos. Para ello, se tuvieron en cuenta los siguientes aspectos:

— Que se trata de una asignatura de 6 créditos (1 crédito = 25 horas de trabajo del alumno) implementada en un Plan de Estudios y en un Departamento, por lo que se procuró que los contenidos se adecuaran al número de créditos, no interfirieran con los de otras asignaturas y que, a su vez, fueran los suficientes y necesarios para la buena marcha de las demás asignaturas de la carrera.

— Que hubiera un equilibrio entre lo que se le iba a exigir al alumno y la información que poseía a principios de curso.

— El contexto científico de la materia y, fundamentalmente, las competencias específicas y transversales que la sociedad va a demandar a los profesionales que formemos.

Entre las muchas preguntas que nos planteamos al iniciar nuestro trabajo estaba la de qué objetivos pretendíamos conseguir; pues bien, a lo largo de los muchos años que llevamos como docentes hemos podido aprender de nuestros errores y, de esta manera, podemos plantearnos unos objetivos fácilmente alcanzables. A grandes rasgos, y tal y como aparece recogido en el Libro Blanco del Grado de Psicología, el objetivo general del título de grado es *«formar profesionales con los conocimientos científicos necesarios para comprender, interpretar, analizar y explicar el comportamiento humano, y con las destrezas y habilidades básicas para evaluar e intervenir en el ámbito individual y social a lo largo del ciclo vital, con el fin de promover y mejorar la salud y la calidad de vida»*. Pues bien, en este objetivo general hay una parte fundamental que corresponde a la Psicometría, la relacionada con la medición y cuantificación de las variables psicológicas para poder llevar a cabo el proceso de evaluación; para ello deberá incluir entre sus contenidos todos aquellos que nos permitan la elaboración de los instrumentos científicos para llevarla a cabo. Teniendo esto en cuenta, hemos considerado que nuestros alumnos deberán adquirir los conocimientos necesarios acerca de:

— Los fundamentos de la Teoría de la Medición.

— Las principales teorías y modelos para la construcción de tests y escalas.

— Los conceptos de fiabilidad y validez, así como de sus distintas formas de obtención e interpretación.

— Las distintas formas de asignación e interpretación de las puntuaciones obtenidas por los sujetos en los tests.

En definitiva, que lleguen a tener un conocimiento suficiente de lo que es la Psicometría y de los métodos y técnicas que aporta a la Psicología científica en general y a la evaluación psicológica en particular.

Al final del libro se incluyen un glosario en el que se recogen, ordenados alfabéticamente, los principales conceptos que han ido apareciendo a lo largo de los distintos temas y las tablas estadísticas necesarias.

Si partimos de que los principios metodológicos del EEES requieren que los materiales didácticos contribuyan a facilitar:

— el aprendizaje autónomo

— el aprendizaje orientado a la adquisición de competencias genéricas y específicas que implica no sólo la adquisición de conocimientos, sino también el desarrollo de habilidades y destrezas

— la evaluación continua de los aprendizajes y

— el seguimiento y tutorización del proceso de aprendizaje

Además de este libro básico, los alumnos dispondrán de los siguientes materiales: una guía de estudio que les facilite el trabajo; un formulario con las tablas estadísticas y un libro de problemas resueltos de Psicometría. Asimismo, podrán acceder al curso virtual de la asignatura en el que se irán incluyendo todas las orientaciones y actualizaciones que se considere oportuno.

Mª Isabel Barbero García

(Madrid, Septiembre de 2015)

TEMA 1

INTRODUCCIÓN A LA PSICOMETRÍA

María Isabel Barbero García

SUMARIO

1. ORIENTACIONES DIDÁCTICAS

En este primer tema se intenta dar una visión global de la Psicometría para que los alumnos se familiaricen con la disciplina cuyo estudio van a abordar; para que conozcan no sólo lo que es la Psicometría y cuál ha sido su evolución y desarrollo a lo largo de los años, sino para que comprendan el papel que desempeña en el proceso de investigación científica.

Para ello, después de hacer una revisión del concepto de Psicometría, y ofrecer una definición para mayor clarificación del mismo, se hace un análisis del área de conocimiento denominada *Metodología de las Ciencias del Comportamiento* en la que está incluida la Psicometría y que coincide con el nombre del Departamento. Este análisis va a permitir a los alumnos comprender la importancia que tienen algunas de las asignaturas que han de estudiar en el Grado de Psicología como son: *Introducción al Análisis de datos; Fundamentos de Investigación; Diseño y Análisis de datos* y, finalmente, *Psicometría*, puesto que les van a proporcionar las claves para poder llevar a cabo cualquier proceso de investigación científica.

Partiendo de que el concepto de medición es algo intrínseco a la Psicometría, y aceptando la necesidad y posibilidad de llevar a cabo mediciones en el marco de la Psicología como ciencia positiva que es, en este tema se hace un análisis de los dos caminos a través de los cuales se fue desarrollando la Psicometría: *Los Estudios de Psicofísica y los Estudios de las Diferencias Individuales*. Los primeros dieron lugar a los *Métodos de Escalamiento de estímulos* y los segundos al *Método de los Tests* para el escalamiento de los sujetos. En un principio estas dos vías de desarrollo siguieron caminos muy diferenciados pero, hoy día, esta separación está superada y suele mantenerse únicamente por motivos didácticos.

Teniendo en cuenta que la asignatura de Psicometría es una asignatura cuatrimestral dentro del plan de estudios de nuestra Universidad, ha sido necesario seleccionar los temas a incluir en este libro dada la amplitud de contenidos que abarca nuestra disciplina, y se ha creído conveniente centrarnos fundamentalmente en el estudio de la Teoría de los Tests. Por eso, no se hace una revisión

exhaustiva de los distintos métodos desarrollados para el escalamiento de los estímulos, sino un breve apunte, y sí una revisión más extensa del origen y desarrollo de los tests como instrumentos de medida y de las distintas teorías de los tests, para poder abordar en los temas que siguen el problema de su construcción, evaluación y aplicación.

Una vez estudiado este tema, los alumnos deberán tener muy claro el importante papel que juega la Psicometría en el marco de la Psicología científica. Deberán conocer los caminos a través de los que se fue desarrollando nuestra disciplina, sabiendo diferenciar entre los distintos métodos de escalamiento según que el objetivo perseguido sea el escalamiento de los estímulos, el de los sujetos o el de ambos (escalamiento de las respuestas).

Deberán aprender a conocer y valorar a figuras tan importantes como Fechner, Stevens y Thurstone por sus aportaciones para la elaboración de escalas psicofísicas y psicológicas y a figuras como Galton, Pearson, Cattell, Binet, Terman y tantos otros, por sus trabajos pioneros en el campo de la medida de las diferencias individuales, trabajos que dieron lugar al desarrollo del método de los tests y sentaron las bases para su evolución hasta alcanzar las cotas de desarrollo que tienen hoy en día, por ejemplo, los tests adaptativos informatizados.

Otro de nuestros objetivos es que los alumnos entiendan que los tests, como cualquier otro instrumento de medición, son imperfectos, y que las puntuaciones que obtienen los sujetos cuando se les aplican no representan con exactitud su verdadera puntuación en aquello que se está midiendo ya que están afectadas de errores de medida cuya cuantía es necesario estimar.

Deberán conocer las distintas teorías de los tests que se han ido desarrollando a lo largo de los años y que incluyen una serie de modelos matemáticos (o funciones). Estos modelos permiten establecer una relación entre las puntuaciones empíricas que obtienen los sujetos en los tests y su verdadera puntuación (o su nivel real) en la característica que se desea medir, y hacer estimaciones acerca de los errores presentes en el proceso de medición. Las diferentes funciones han dado lugar a diferentes teorías de los tests siendo las dos más importantes la Teoría Clásica de los Tests (TCT) y la Teoría de Respuesta al Ítem (TRI). Se puede hacer referencia también a la Teoría de la Generalizabilidad (TG) que, a pesar de que hoy en día no tiene la relevancia que se le auguraba en un principio, supuso un esfuerzo para tratar de dar solución a algunos de los problemas que quedaban sin resolver en la TCT.

2. APROXIMACIÓN AL CONCEPTO DE PSICOMETRÍA

Antes de comenzar el estudio de cualquier disciplina es necesario tener un conocimiento claro de qué es lo que se va a estudiar y porqué. Por lo tanto, dado que los contenidos se refieren a la Psicometría, el primer paso que hemos de dar es conceptuar el término y explicar lo que se entiende por Psicometría, teniendo en cuenta, además, que no se trata de un concepto estático sino algo dinámico que irá evolucionando gracias a las aportaciones de las investigaciones realizadas en su campo, y se ampliará en la medida en que se vayan incorporando nuevos conocimientos.

La aproximación al concepto de la Psicometría, como al de cualquier otra disciplina, es una tarea difícil dada la variedad de facetas que presenta; sin embargo, si se quiere hacer un estudio riguroso del mismo puede abordarse a través de distintos caminos. Cada uno de estos caminos ofrecerá una información parcial; por eso, en la medida en que se utilice un mayor número de ellos se dispondrá de una información más completa.

Uno de los caminos a seguir para aproximarnos al concepto de Psicometría, y quizás el más inmediato, puede ser el análisis etimológico del término formado por las palabras griegas *«Psykhe» y «Metrum»*, que literalmente significa *«Medida de la Psykhe»*. Sin embargo, dado que la información que ofrece esta vía es bastante amplia y ambigua es necesario completarla; por eso se ha hecho un estudio de las definiciones dadas por algunos expertos (Cerdá, 1970; Cliff, 1979; García-Cueto, 1993; Macià, 1982; Martínez Arias, 1995; Melia, 1990; Morales, 1975; Muñiz, 1998; Musso, 1970; Nunnally, 1973; Rivas, 1976; Santisteban, 1990; Seoane, 1980 y Yela, 1968, entre otros) y, mediante un análisis de su contenido, se puede concluir que:

> *La Psicometría es una disciplina metodológica, dentro del área de la Psicología, cuya tarea fundamental es la medición o cuantificación de las variables psicológicas con todas las implicaciones que ello conlleva, tanto teóricas (posibilidades y criterios de medición) como prácticas (cómo y con qué se mide).*

Analizando la definición anterior, se puede decir que la Psicometría deberá ocuparse en primer lugar de la *justificación y legitimación de la medición psicológica*, para lo cual deberá: *a)* desarrollar modelos formales que permitan representar los fenómenos que se quieren estudiar y posibiliten la transformación de los hechos en datos, *b)* validar los modelos desarrollados para determinar en qué medida representan la realidad que pretenden y *c)* establecer las condiciones que permitan llevar a cabo el proceso de medición. En segundo lugar, deberá también ocuparse de las *implicaciones prácticas y aplicadas que dicha medición conlleva*, proporcionando los métodos necesarios que indiquen, en cada caso concreto, cómo se debe llevar a cabo la cuantificación, y construyendo los instrumentos necesarios y adecuados para poder efectuarla. Esta vertiente aplicada de la Psicometría, referida a la construcción y evaluación de los instrumentos de medición, que es de la que nos vamos a ocupar fundamentalmente en este libro, no debe confundirse con el uso que se haga de los instrumentos una vez construidos. Un instrumento puede estar bien construido y, sin embargo, ser utilizado de forma incorrecta.

Si se quiere medir de alguna manera la extraversión, será necesario desarrollar el instrumento científico adecuado para llevar a cabo el proceso de medición, esa es la parte que le incumbe al

psicómetra; otra cosa muy diferente es el uso, bueno o malo, que se haga del instrumento una vez construido.

La importancia de la Psicometría, como disciplina a la que incumbe todo aquello relacionado con la medición de variables psicológicas, se justifica si se tiene en cuenta que en Psicología, como en las demás ciencias empíricas, el objetivo final es la *descripción, explicación y predicción* de los fenómenos de interés (en nuestro caso los fenómenos psicológicos) y se podrá cumplir mejor dicho objetivo mediante el proceso de medición. Por eso, aunque la Psicometría no tiene un campo de aplicación específico como sucede con otras disciplinas, su campo de aplicación abarca todos los campos de la Psicología: personalidad, procesos cognitivos, actitudes, etc., y juega un papel importantísimo ya que contribuye a fundamentar, elaborar y contrastar todas las teorías psicológicas. Es dentro de este marco donde se justifica la medición.

Ahora bien, las mediciones llevadas a cabo sin un contexto teórico o aplicación práctica que les sirva de base rara vez justifican el tiempo y el dinero que se invierte en ellas. Es necesario evitar caer en la tentación de considerar que la medición es la piedra de toque de la respetabilidad científica; como señala Miller (1982):

> *«...muchos psicólogos se han precipitado a buscar números antes de saber lo que esos números pueden significar... Todavía se pueden encontrar psicólogos que llevan a cabo mediciones de gran complicación y exageradamente precisas sólo para demostrar hasta qué punto el psicólogo puede llegar a ser científico. Hay gente que no admite que, si apenas merece la pena hacer una cosa, hacerla bien sigue sin merecer la pena»* (pág. 115).

3. LA PSICOMETRÍA EN EL MARCO DE LA METODOLOGÍA DE LA PSICOLOGÍA COMO CIENCIA DEL COMPORTAMIENTO

En el apartado anterior se definió la Psicometría como una disciplina metodológica dentro del área de la Psicología; por eso, a la hora de adscribir las asignaturas del plan de estudio a un Departamento, la Psicometría lo fue al de *Metodología de las Ciencias del Comportamiento*.

Si se hace un análisis de esta denominación, nos encontramos con dos conceptos fundamentales: el concepto de *Metodología* y el de *Ciencias del Comportamiento*. En la medida en que puedan ser aclarados, se tendrá un mejor conocimiento del marco en el que se sitúa la Psicometría.

Partiendo de un análisis etimológico del término, *Metodología* significa *Tratado de los Métodos*, y teniendo en cuenta que dentro del marco de la ciencia el término *Método* hace referencia al camino que se debe seguir para poder conseguir el objetivo de la ciencia, se puede considerar que:

> *...la Metodología estudia las estrategias y procedimientos que, de una forma más o menos estructurada, se utilizan para la obtención de los conocimientos que constituyen una disciplina científica.*

Por otra parte,

> *...las Ciencias del Comportamiento son aquellas que estudian la «conducta» mediante la utilización del método científico, con el fin de encontrar estructuras generales o leyes.*

En este contexto el término «conducta» se utiliza en sentido amplio, y hace referencia a la actividad de un organismo ante una situación concreta que estará determinada biológica y socialmente.

Teniendo en cuenta que el método científico es el método común a todas las ciencias, que proporciona un marco general a cualquier proceso de investigación científica, pero que puede adaptarse a las peculiaridades de cada una de ellas en función de sus problemas específicos y de su objeto de estudio:

> *La Metodología de las Ciencias del Comportamiento estará referida al estudio del método general de la ciencia y de las estrategias o métodos específicos que deberán desarrollar cada una de ellas, en función de sus peculiaridades, para poder llevar a cabo su tarea.*

La Psicología, como ciencia del comportamiento, tiene su propio objeto de estudio y sus propios problemas; por eso, deberá adaptar el método general de la ciencia, el método científico, al marco concreto de cada uno de los problemas, desarrollando las estrategias y técnicas complementarias que le permitan abordar su estudio y tratar de darles solución. A estas técnicas algunos autores las denominan *Técnicas metódicas* (Cruz Hernández, 1976; Fernández Ballesteros, 1980; Fernández-Trespalacios, 1979 y Moreno, 1983 entre otros).

Dentro del marco de la Metodología de las Ciencias del Comportamiento, y más concretamente de la Metodología de la Psicología como ciencia del comportamiento que es, encontramos una gran cantidad de contenidos relacionados entre sí por su carácter procedimental. Aunque estos contenidos por razones didácticas, y para que los alumnos aprendan a relacionar los contenidos entre sí, se hayan articulado en cuatro asignaturas dentro del Grado de Psicología, en realidad se pueden considerar tres grandes bloques cuyos orígenes van unidos a los de la Psicología científica: *Diseños de investigación, Análisis de datos y Psicometría*.

El bloque dedicado a los *Diseños de investigación* atiende, fundamentalmente, a aquella faceta de la investigación científica cuya tarea fundamental es la operativización de las variables de la hipótesis y la elaboración de un plan de trabajo, o procedimiento para la recogida de los datos, que sea coherente con la hipótesis; puesto que, de acuerdo con Arnau (1990) el concepto de diseño

de investigación *está esencialmente vinculado a la elección y especificación del procedimiento para la obtención de los datos relevantes a la hipótesis* (pág. 13). Para llevar a cabo esta tarea, es necesario analizar, entre otros, los siguientes aspectos: la naturaleza de las variables, sus posibilidades de manipulación, la elección de aquellas que sean de interés para la investigación objeto de estudio, la detección de variables extrañas y formas de control, los criterios de selección y asignación de unidades de observación, la estructuración interna del procedimiento con delimitación de situaciones, tareas, etc. (Sarriá, 1991). En la medida en que se resuelvan todos estos aspectos se reducirá la incertidumbre que conlleva la elección de un diseño de investigación.

El bloque correspondiente al *Análisis de datos* será el encargado de proporcionar las técnicas necesarias para llevar a cabo el tratamiento estadístico de los mismos, tratamiento que puede ir desde la simple descripción o representación gráfica a procedimientos más complejos de ajuste de modelos o contrastes de hipótesis.

Por último, la *Psicometría*, comparte con el resto de las disciplinas psicológicas tanto el objeto de estudio: la conducta humana, como el método: el método científico; entonces, lo que realmente la caracteriza es la peculiar adaptación que hace de éste al objeto de estudio de la Psicología. Al incluir todo la referente a la medición, la Psicometría proporciona las reglas que van a permitir llevar a cabo el proceso de operativización de las variables que se quieren medir. Una vez obtenidas las medidas mediante la asignación de números, los modelos psicométricos permitirán un análisis del error que les afecta (fiabilidad de las medidas) y, finalmente, los estudios de validación permitirán hacer inferencias acerca de las relaciones entre los datos empíricos obtenidos (medidas) y el constructo o variable psicológica que se quiere medir.

Como se desprende de todo lo anterior, a pesar de que estos tres bloques tienen una entidad propia, cada uno tiene una enorme influencia sobre los demás, y juntos interactúan en el proceso global e integrado que es la investigación científica.

> **Nota:** Creo que esta pequeña introducción permitirá a los alumnos conocer la relación que existe entre las asignaturas del Departamento de Metodología de las Ciencias del Comportamiento y su importancia dentro de los estudios de Grado de Psicología.

4. LA MEDICIÓN PSICOLÓGICA

Dado que en las asignaturas de *Introducción al Análisis de Datos* y *Fundamentos de Investigación* ya se abordó el tema de la medición en Psicología no vamos a extendernos nosotros en su exposición, pero sí queremos resaltar la importancia del mismo dado que se trata de un problema inherente al desarrollo de la Psicometría y de la Psicología científica. Hasta el momento en que se acepta la posibilidad de medir lo psicológico no se consideraba que la Psicología fuera una ciencia.

De acuerdo con Coombs, Dawes y Tversky (1981), consideramos que uno de los papeles fundamentales asignados a la Ciencia es la descripción, explicación y predicción de los fenómenos observables por medio de unas cuantas leyes generales que expresen las relaciones entre las propiedades de los objetos investigados. En las Ciencias más avanzadas las leyes expresan relaciones cuantitativas, lo cual indica que las propiedades de los objetos se pueden representar por medio de números mediante un proceso de *medición*. la Psicología como Ciencia tendrá su base científica en la medición, que le permitirá contrastar empíricamente las hipótesis planteadas.

Para Nunnally (1,970), la medición se reduce a algo muy sencillo, *consiste en un conjunto de normas para asignar números a los objetos de modo tal que estos números representen cantidades de atributos* (pág. 23*), entendiendo por atributos las características de los objetos y no los objetos mismos*. Cuando decimos que queremos medir una mesa, lo que realmente queremos hacer es medir alguna característica de la mesa, como por ejemplo su longitud o su altura, no la mesa como objeto. Esto implica que la medición conlleva un proceso de abstracción. Hoy día la medición se toma, en general, como la asignación de números a entidades, acontecimientos o sucesos, con el fin de representar sus propiedades y sus relaciones.

Ya en la antigüedad, cuando se quería medir algún atributo físico de los objetos, como podía ser su peso, su longitud, etc., se trataban de desarrollar los instrumentos pertinentes y nadie dudaba acerca de la posibilidad de llevar a cabo tales mediciones. Sin embargo, la polémica surgió cuando en lugar de querer medir atributos físicos se intentaron medir atributos psicológicos, puesto que existían serias dudas acerca de la posibilidad de llevar a cabo tales mediciones. Una de las formas de paliar esas dudas era mostrar que las variables psicológicas se podían cuantificar y que los procedimientos que se utilizaban para ello permitían establecer relaciones cuantitativas entre las variables.

No obstante, a nadie se le escapa la dificultad que entrañaba la medición de características psicológicas dada la singularidad de las mismas y, por lo tanto, las dificultades que hubo que ir superando hasta que se consiguió que se aceptara la necesidad y posibilidad de medir este tipo de variables. La dificultad principal derivaba, fundamentalmente, de que a diferencia de los atributos físicos de los sujetos, como el peso y la estatura, que pueden ser medidos directamente con los instrumentos pertinentes, la mayoría de los atributos psicológicos como por ejemplo la inteligencia, el autoritarismo y la introversión, son conceptos abstractos, denominados también *constructos teóricos (variables latentes)*, cuya medición no puede llevarse a cabo de forma directa sino que debe inferirse a través de la medición de una serie de conductas representativas de dicho constructo. En este sentido, Zeller y Carmines (1980) plantearon una nueva concepción de la medición; consideraron que se trataba de *un proceso mediante el cual se enlazan conceptos abstractos (los constructos inobservables directamente), con indicadores empíricos observables directamente (las conductas).*

Si por ejemplo se quiere medir la *inteligencia,* lo primero que se nos ocurre preguntar es ¿qué es la inteligencia? Entonces nos daremos cuenta de que es algo abstracto, inobservable y que, por lo tanto, no se puede medir directamente. Sin embargo, estamos acostumbrados a oír decir de las personas que son poco o muy inteligentes lo cual quiere decir que, de alguna manera, se puede evaluar eso que nosotros llamamos inteligencia. Una forma de hacerlo puede ser analizar cómo se comportarían las personas, a las que consideramos inteligentes, ante determinadas situaciones y, posteriormente, crear el instrumento adecuado para medir esas conductas. De esta manera, la variable psicológica *inteligencia,* que es un constructo teórico inobservable de forma directa, se puede manifestar a través de una serie de conductas que ya sí son observables directamente y, por lo tanto, pueden ser medidas mediante el instrumento adecuado. A partir de los resultados obtenidos en el proceso de medición, se podrán hacer inferencias acerca del grado en que cada uno de los sujetos evaluados posee el constructo de interés, en nuestro caso, del grado de inteligencia que poseen.

Este tipo de medición se suele denominar *medición por indicadores* entendiendo que, dado que las variables psicológicas no se pueden medir de forma directa, es necesario seleccionar una serie de indicadores que sí pueden ser medidos directamente, y que se supone están en estrecha relación con el constructo o variable psicológica que se quiere medir.

Hoy día se han desarrollado distintos procedimientos para la medida de las sensaciones, aptitudes, actitudes, etc., pero hasta hace poco no tenían la fundamentación lógica necesaria para su justificación. Como señala Muñiz (1992), los únicos métodos que había para evaluar la calidad métrica de las medidas obtenidas eran una colección de técnicas estadísticas encuadradas bajo las denominaciones de *fiabilidad y validez;* de ahí que no era de extrañar que las mediciones llevadas a cabo en el campo de la Psicología fueran consideradas, tanto cualitativamente como cuantitativamente, de orden inferior a las realizadas en el campo de la física. Sin embargo, los desarrollos recientes han demostrado que, aunque las medidas obtenidas al medir variables psicológicas puedan ser menos precisas que las realizadas en el campo de la Física el estatus teórico de las mediciones puede justificarse al mismo nivel (pág. 232).

Además del problema anterior, a la hora de medir variables psicológicas nos encontramos con otro problema, el de las unidades de medida que se van a utilizar al interpretar las puntuaciones obtenidas por los sujetos. Si nosotros medimos la longitud de una mesa el resultado lo podemos expresar en centímetros, si lo que evaluamos es el peso de cualquier objeto el resultado se expresará en gramos o kilogramos, pero si lo que medimos es la inteligencia de un niño o su capacidad para las matemáticas, ¿qué unidades de medida podemos utilizar para dar un significado a los resultados obtenidos?

En Psicología, hay dos formas fundamentales de abordar el problema, una *referida a normas* y otra *referida al criterio.* La forma más habitual de proceder es la primera, la referida a normas, que consiste en comparar los resultados obtenidos por ese niño con los obtenidos por un grupo de niños que forman el llamado *grupo normativo* y que pertenecen a su misma población; es decir, que pertenecen a su misma clase, son de su misma edad, etc. En otras ocasiones la interpretación se hace en relación a un criterio previamente establecido; los resultados obtenidos se comparan con ese *criterio* (un punto crítico) y la superación o no del mismo es lo que va a dar significado a las puntuaciones obtenidas.

Las dificultades que entraña la medición psicológica se comprenden mejor si, como señala Muñiz (1998), se tiene en cuenta que la conducta humana se desarrolla en una banda acotada por una base neurobiológica y un entorno sociocultural y surge, por lo tanto, de la interacción entre nuestra constitución biológica y la estimulación ambiental.

> **Nota:** El lector interesado en profundizar en el problema de la medición en Psicología puede consultar, entre otros, el libro de José Luis Meliá (1990) *Introducción a la medición* y el de Joel Michell (1999), *Measurement in Psychology: a critical history of a methodological concept.*

Sea cual sea el campo de aplicación de la medición psicológica (procesos básicos, personalidad, procesos cognitivos, actitudes, valores, etc.) hay una serie de objetivos comunes fundamentales: en primer lugar estimar los errores aleatorios que conlleva toda medición (fiabilidad de las medidas) y garantizar que la misma no es algo inútil sino que sirve para explicar y predecir los fenómenos de interés (validez de las medidas). Todos estos aspectos se irán analizando a lo largo de los temas siguientes, después de haber hablado acerca de los orígenes y vías de desarrollo de la Psicometría, o lo que es lo mismo de los orígenes y vías de desarrollo de la medición psicológica.

5. ORÍGENES Y VÍAS DE DESARROLLO DE LA PSICOMETRÍA

Partiendo de que es necesaria y posible la medición en Psicología, y teniendo en cuenta que la Psicometría es la disciplina que entre sus contenidos incluye todo lo relativo a la medición, vamos a ir analizando brevemente los hechos fundamentales que contribuyeron a su desarrollo y convirtieron a la Psicología en una ciencia positiva independiente de la Filosofía.

De acuerdo con Yela (1968), podemos decir que existieron dos motivos fundamentales que posibilitaron la introducción de la medición en Psicología. Uno de ellos, la tendencia a formular los problemas científicos en términos matemáticos; el otro motivo, el enfrentamiento de la Psicología, hacia la mitad del siglo XIX con dos problemas fundamentales: el primero, el estudio cuantitativo de las relaciones entre las características físicas de los estímulos y las sensaciones que suscitan en los sujetos en función de las cuales se asignan valores numéricos a los estímulos: *el problema psi-*

cofísico; el segundo, el problema de *la cuantificación de las diferencias individuales* mediante la asignación de números a los sujetos en función del grado en que manifiesten un atributo o conducta.

Así pues, se puede situar el origen de la Psicometría hacia la mitad del siglo XIX y, a partir de ese momento, se va a desarrollar, fundamentalmente, a través de estas dos vías:

a) los estudios de Psicofísica que dieron lugar al desarrollo de modelos que permitieron asignar valores numéricos a los estímulos y, por lo tanto, que permitieron *el escalamiento de estímulos*.

b) los estudios acerca de las diferencias individuales que dieron lugar al desarrollo de los Tests y de las distintas Teorías de los Tests (modelos), que posibilitaron la asignación de valores numéricos a los sujetos y, por lo tanto, el *escalamiento de los sujetos*.

Estos dos puntos de arranque de la Psicometría (los estudios de psicofísica y la cuantificación de las diferencias individuales) dieron lugar al desarrollo de las dos corrientes que mayor incidencia han tenido en la investigación psicológica: la experimentalista y la correlacional.

En la Psicometría clásica se reservaba el término *Escalamiento* para designar el proceso de construcción de escalas para la cuantificación de estímulos; mientras que todo lo relativo a la cuantificación de los sujetos, que se estudiaba en la *Teoría de los Tests*, no se consideraba parte de dicho proceso. En la actualidad, la distinción entre *Escalamiento y Teoría de los Tests,* aunque se mantiene por cuestiones didácticas, está superada gracias al desarrollo de nuevas teorías de la medición y de numerosas técnicas estadísticas comunes a ambas vertientes de la Psicometría.

6. LOS MÉTODOS DE ESCALAMIENTO

Podemos considerar el escalamiento como el campo de la Psicometría cuyo objetivo fundamental es la construcción de escalas de medida; es decir, la construcción de instrumentos que permitan llevar a cabo mediciones para representar las propiedades de los objetos (estímulos, sujetos o respuestas) por medio de números, de acuerdo a unas normas o reglas.

Dentro del marco de la Psicometría, hay dos supuestos básicos en todos los métodos de escalamiento:

— La existencia de un continuo latente o subyacente, a lo largo del cual varían los objetos psicológicos que se van a escalar (estímulos, sujetos, o respuestas), que no puede ser observado de forma directa.

— Que los objetos psicológicos (estímulos, sujetos o respuestas) pueden situarse de forma ordenada a lo largo de ese continuo.

Por ejemplo, si el atributo o característica que se quiere medir es la inteligencia, supondremos que ese atributo se puede representar a lo largo de un continuo (una recta en términos geométricos), sobre el cual se podrán situar los sujetos de forma ordenada en función del grado de inteligencia que manifiesten (siguiendo con el símil de la representación geométrica, los sujetos se podrían representar a lo largo de la recta por medio de puntos).

Hemos estado considerando como objetos psicológicos a los estímulos, a los sujetos y a las respuestas. Ahora bien ¿qué diferencias nos vamos a encontrar en función de que los objetos psicológicos a escalar sean estímulos, sujetos o respuestas?

Como señalaron Ghiselli, Campbell y Zedeck (1981, pág. 392) *los estímulos son las cosas que los investigadores presentan normalmente a un sujeto con el propósito de elicitar una respuesta*. Partiendo de esta definición, el término estímulo tiene un sentido muy amplio ya que pueden ser considerados como tales no sólo un conjunto de objetos físicos, sino una lista de adjetivos que hagan referencia a algún rasgo de personalidad, una serie de frases que se refieran a alguna variable de actitud, un grupo de personas a las que se vaya a calificar, una serie de problemas de Psicometría, etc.

En cualquier caso, el propósito del escalamiento de estímulos es determinar las características que los sujetos perciben en ellos y, por lo tanto, la respuesta del sujeto (o sujetos) ante la presentación de los estímulos es una respuesta subjetiva que nos va a permitir diferenciarlos y escalarlos; es decir, asignar un valor numérico a cada uno de los estímulos. Cuando se utilizan varios sujetos para el escalamiento de estímulos, el valor que se asigna a cada uno de ellos suele venir determinado por algún índice de tendencia central, la media o la mediana, obtenido a partir de los valores asignados por cada uno de los sujetos a cada estímulo. En este caso serían los estímulos los que se situarían a lo largo del continuo en función de sus valores escalares (los valores asignados) y los sujetos que han emitido juicios acerca de esos estímulos, los que habrían actuado como instrumentos de medida.

Cuando el objeto a escalar son los sujetos se utiliza una muestra, generalmente extraída de forma aleatoria de una población, y todos los sujetos que la componen responden al mismo conjunto de estímulos (los elementos de un test, por ejemplo); posteriormente, las respuestas emitidas por los sujetos ante la presentación de cada uno de los estímulos serán combinadas de alguna manera para proporcionar una puntuación numérica para cada sujeto de la muestra. Las diferencias encontradas en las puntuaciones obtenidas reflejarán las diferencias entre los sujetos respecto al atributo o característica que se está midiendo. Se asume que los estímulos presentados son interpretados de la misma forma por todos los sujetos de la muestra y, por lo tanto, las variaciones encontradas son debidas a las diferencias entre los sujetos. Serán los sujetos los que se situarán a lo largo del continuo que representa la característica que se está midiendo, y los estímulos los que habrán actuado como instrumento de medida.

Hay veces que lo que interesa es situar sobre el continuo, a lo largo del cual varía el atributo que se está estudiando, tanto a los estímulos como a los sujetos. En este caso, las variaciones en-

contradas en las respuestas de los sujetos ante la presentación de los estímulos se atribuyen, no sólo a las diferencias que hay entre los estímulos respecto al grado de atributo que llevan implícito, sino también a las diferencias que hay entre los sujetos en cuanto a su actitud personal frente a cada uno de los estímulos. En estos métodos la posición de cada sujeto respecto del atributo que se está analizando, su actitud, aptitud, sentimientos, etc., es un factor que está determinando su respuesta. Esta tercera aproximación a los métodos de escalamiento fue denominada por Torgerson (1958) *aproximación centrada en la respuesta*.

Vemos, por lo tanto, que los métodos de escalamiento hacen referencia tanto al escalamiento de estímulos como al de sujetos, o al de ambos a la vez.

7. DESARROLLO DE LOS MÉTODOS DE ESCALAMIENTO DE ESTÍMULOS

Dado que en sus orígenes el escalamiento estuvo asociado al escalamiento de estímulos vamos a mantenerlo aquí de esa forma por razones didácticas, tal y como ya se ha comentado, y abordaremos el estudio de todo lo referente al escalamiento de sujetos dentro del marco de la teoría de los tests.

El origen de los métodos de escalamiento de estímulos tuvo lugar cuando la Psicología se enfrentó con el problema de cuantificar, de alguna manera, las relaciones existentes entre las características físicas de los estímulos y las sensaciones que dichos estímulos suscitan en los sujetos; es decir, cuando la Psicología se enfrenta con el problema psicofísico.

7.1. Métodos de escalamiento psicofísico

A principios del siglo XIX el filósofo y pedagogo alemán Herbart acuñó el concepto de *umbral mínimo* para designar a la mínima intensidad que tiene que tener un estímulo para que se pueda percibir; vemos ya, por lo tanto, un primer intento de relacionar la intensidad de los estímulos con las sensaciones que producen. A partir del concepto de umbral mínimo Weber (1795-1878) comenzó a desarrollar una serie de procedimientos experimentales, *los métodos psicofísicos,* que iban a permitir el cálculo de los umbrales y desarrolló su famosa *Ley de Weber*, que establece que:

> *... el incremento de magnitud que debe experimentar un estímulo (ΔE) para que el sujeto perciba que se ha producido un cambio, es una proporción constante de su magnitud inicial (E):*

$$K = \frac{\Delta E}{E} \qquad [1.1]$$

A la constante K se la conoce como constante de Weber; el (ΔE) es el incremento mínimo que ha de experimentar la magnitud de un estímulo (magnitud física) respecto a la magnitud inicial (E) para que el sujeto perciba un cambio mínimo en la sensación (magnitud psicológica), a este cambio mínimo de sensación le denominó Weber *diferencia apenas perceptible* (dap).

Aunque a Weber se deben los primeros intentos de establecer una ley general para formular la noción de umbral, la figura más representativa fue Fechner (1801-1887), que desarrolló una serie de métodos denominados *métodos psicofísicos indirectos* que permitían elaborar unas escalas denominadas *escalas psicofísicas* (Baird y Noma, 1978; Barbero, 1993/1999, 2007; Blanco, 1996; Fechner, 1966; Muñiz, 1991, entre otros).

Supongamos que se quieren escalar una serie de estímulos, que varían en cuanto a su peso, respecto al grado de pesadez que producen en los sujetos. En este caso, cuando a los sujetos se les presenten los distintos estímulos a escalar, deberán emitir un juicio acerca del grado de pesadez que han percibido al sopesar cada uno de ellos. El peso es una característica física de los estímulos que varía a lo largo de un continuo físico y hay instrumentos adecuados para su medición; por el contrario, la pesadez es una característica psicológica o subjetiva que varía a lo largo de un continuo psicológico, y los sujetos actúan como un instrumento de medida asignando valores a cada uno de los estímulos en función del grado de pesadez que les hayan provocado al presentárselos. Estos valores son psicológicos o subjetivos.

Por lo tanto, en el escalamiento psicofísico se tienen dos continuos, un continuo físico a lo largo del cual varían los estímulos y uno psicológico a lo largo del cual variarán las sensaciones que dichos estímulos producen en los sujetos. Para construir la escala psicofísica, será necesario ver qué relación funcional se puede establecer entre los dos continuos.

Según Fechner, la función que mejor representa la relación entre los dos continuos es una *función logarítmica* cuya fórmula es:

$$S = C \ln E + A \qquad [1.2]$$

Siendo:

S = valor en la escala de sensación.

E = valor del estímulo.

C y A = la pendiente y la ordenada en el origen de la función logarítmica.

Para poder medir la relación entre la magnitud de los estímulos y las sensaciones que producen Fechner introdujo los conceptos de *umbral absoluto* y *umbral diferencial*. Para Fechner, *el umbral absoluto* sería la magnitud física del estímulo que se requiere para que se produzca una sensación y el *umbral diferencial* sería el incremento mínimo, en la magnitud física del estímulo, que se requiere para que el sujeto perciba un cambio de sensación.

La ley de Fechner establece que cuando la magnitud física del estímulo está en el umbral absoluto la sensación es nula, en ese punto se establece el origen en la escala de sensación (será el 0 de la escala de sensación), y que si se aumenta la estimulación en proporción geométrica las sensaciones aumentarán aritméticamente. Es decir, que cada vez se necesitará un mayor incremento en la estimulación física para que se perciba un cambio en la sensación.

Los principales supuestos en los que se basa la ley de Fechner son los siguientes:

1. Asume la ley de Weber:

$$K = \frac{\Delta E}{E}$$

   Es decir, que el incremento de magnitud que debe experimentar un estímulo (ΔE) para que el sujeto perciba que se ha producido un cambio es una proporción constante de su magnitud inicial (E).

2. Asume que todas las diferencias apenas perceptibles (d.a.p) son psicológicamente iguales.

   Esto implica que siempre que se produzca un aumento en la magnitud del estímulo igual a un umbral diferencial (ΔE), sea cual sea el valor de esa magnitud, la sensación aumentará siempre en la misma cuantía. Esto puede expresarse así: $\Delta S = I$, donde I es una constante.

3. Establece que el origen de la escala de sensación, es decir el punto cero de la misma, corresponde al umbral absoluto, es decir al valor del estímulo correspondiente al valor absoluto en la escala física.

EJEMPLO:

Supongamos que se quiere medir la capacidad auditiva de una persona; para ello, se le presentan una serie de estímulos de distinta intensidad comenzando por uno cuya intensidad de sonido esté por debajo de la capacidad auditiva de la persona. Poco a poco se va aumentando la intensidad del sonido hasta que la persona empiece a detectar el estímulo. En ese momento ha habido un cambio, la persona ha pasado de no tener ninguna sensación a percibir el sonido. El valor del estímulo que ha provocado ese cambio corresponde al *umbral absoluto* de la persona y marca el origen de la escala de sensación (el punto cero). Si a partir del valor correspondiente al umbral absoluto se va aumentando muy lentamente la magnitud del estímulo (en nuestro ejemplo la intensidad del sonido), llega un momento en que la persona detecta que ha habido un cambio y se produce en ella un cambio de sensación. En ese momento se ha encontrado su primer *umbral diferencial,* que correspondería al incremento mínimo que tiene que experimentar la intensidad del estímulo para que la persona perciba un cambio de sensación. Al cambio de sensación se le denomina *diferencia apenas perceptible (dap)*, por lo que se dirá que se ha producido una (dap) a partir del umbral absoluto.

Siguiendo con el mismo razonamiento se irían calculando los diferentes umbrales diferenciales que permitirán obtener dos series de valores, una correspondiente a los distintos valores del estímulo (escala física), y la otra serie la escala de sensación (escala psicológica) cuyos valores se obtendrán a base de ir sumando las *dap* que se han ido produciendo a partir del origen. Una vez obtenidas las dos series de valores sería necesario comprobar la relación que existe entre ellas y si esta relación es logarítmica como postula Fechner.

Si representamos gráficamente sobre unos ejes de coordenadas los pares de valores obtenidos se observará la relación funcional que existe entre ellos y que según Fechner es logarítmica.

Supongamos el caso hipotético de que el umbral absoluto de una persona para el peso son 10 gramos; es decir, que hasta que no se le presente al sujeto un estímulo cuyo peso sean 10 gramos el sujeto no percibe ninguna sensación, supongamos también que para que el sujeto note que hay un cambio en el peso del estímulo se necesita aumentar 2 gramos el peso inicial, lo cual supone que el umbral diferencial de ese sujeto y para ese estímulo inicial es de dos gramos, bajo estos supuestos la constante de Weber sería $K = 0{,}2$ dado que:

$$K = \frac{2}{10} = 0{,}2$$

Si se cumplieran los supuestos de la ley de Fechner obtendríamos los siguientes valores para el estímulo (continuo físico) y la sensación (continuo psicológico):

E	S (dap)
10	0
12	1
14,4	2
17,28	3

¿Cómo se han obtenido estos valores?

Al umbral absoluto le corresponde el cero en la escala de sensación. Si tal y como hemos comentado anteriormente, para que la persona note un cambio de sensación es necesario aumentar el peso del estímulo en 2 gramos, la constante de Weber será 0,2 y el umbral diferencial 2 gramos.

Por lo tanto, cuando el valor del estímulo pasa de 10 a 12 gramos se ha producido en el sujeto 1 (dap), por eso en la escala psicológica aparece el 1.

Sabiendo que la constante de Weber es $K = 0,2$ y que la ley de Fechner asume la ley de Weber, podemos ir averiguando los distintos valores tanto de la escala física como de la psicológica. El siguiente valor será: $0,2 = \frac{\Delta E}{12} \rightarrow \Delta E = 0,2 \times 12 = 2,4$

Cuando el valor del estímulo es de 12 gramos y la constante de Weber 0,2 se necesita un aumento de 2,4 gramos para que el sujeto perciba un cambio en la sensación. Por lo tanto al valor del estímulo igual a 14,4 (12 + 2,4) en la escala física le corresponde el valor de 2 en la escala de sensación puesto que desde el umbral absoluto se han producido 2 dap.

De esta manera se irían obteniendo los valores de ambas escalas suponiendo que los datos se ajustaran a la función de Fechner.

Si se quiere calcular el valor de un estímulo cualquiera, el n-esimo, se debería aplicar la siguiente fórmula:

$$E_n = E_{n-1} + (E_{n-1})K = E_{n-1}(1+K) \qquad [1.3]$$

En el ejemplo que hemos puesto no sería necesario averiguar el ajuste de los datos a la función logarítmica ya que, de antemano, sabemos que el ajuste es perfecto puesto que los hemos ido obteniendo asumiendo los supuestos de la ley de Fechner; ahora bien, el problema que hay que abordar es el de comprobar si, en situaciones reales, se verifica la ley.

La forma de comprobarlo es sencilla, en primer lugar hay que seleccionar los valores del estímulo y, posteriormente, por medio de una serie de métodos llamados métodos psicofísicos se irán obteniendo los datos y se irá elaborando la escala de sensación. Una vez obtenidas las dos escalas lo único que hace falta comprobar es si los datos obtenidos se ajustan o no a la función logarítmica de Fechner.

Ahora bien, el problema fundamental está en la obtención experimental de los umbrales absoluto y diferencial que permitirán medir la capacidad de detección y de discriminación de los sujetos. Para ello, Fechner desarrolló una serie de procedimientos entre los que destacaremos: *el método de los límites, el método de ajuste y el método de los estímulos constantes*. El procedimiento general que se sigue para llevar a cabo el escalamiento (procedimiento que variará en función del método utilizado) consiste en presentar a una muestra de sujetos el conjunto de estímulos a escalar y, en función de las respuestas emitidas por aquellos, a lo largo de los distintos experimentos, se asignará un valor numérico a cada uno de los estímulos.

En el *método de los límites*, también llamado de *los cambios mínimos*, es el experimentador el que va modificando la intensidad del estímulo. Cuando se quiere calcular el umbral absoluto comenzará, o bien por un estímulo de baja intensidad e irá aumentándola hasta que el sujeto que los va a evaluar comience a detectarlo, o bien por uno de alta intensidad e irá reduciéndola poco a poco hasta que el sujeto deje de percibirlo. En el caso de querer calcular los umbrales diferenciales el experimentador seleccionará un valor del estímulo como estándar y, junto a él, irá presentando al sujeto un estímulo de comparación cuya intensidad irá manipulando hasta que el sujeto considere que la intensidad del estímulo de comparación es igual que la del estímulo estándar.

En el *método de ajuste* también llamado de *error promedio*, la diferencia con respecto al método anterior radica en que, en lugar de ser el investigador el que manipula la intensidad de los estímulos, es el propio sujeto el que la va modificando, aumentándola o disminuyéndola, hasta encontrar el valor de sus umbrales.

En el *método de los estímulos constantes* se asume que cuando un estímulo se presenta a un mismo sujeto en repetidas ocasiones no siempre es percibido y que aún cuando lo perciba no siempre le produce la misma sensación. Partiendo de este supuesto, para averiguar el umbral absoluto, cada estímulo se presenta a los sujetos un número elevado de veces en orden aleatorio y, en cada ocasión, cada sujeto ha de decir si lo ha percibido o no. El umbral absoluto es la magnitud del estímulo que ha sido percibido por los sujetos el 50% de las veces que se ha presentado. Para el cálculo del umbral diferencial, al igual que en los métodos anteriores, se fija un valor del estímulo como estándar y, a continuación, se van presentando una serie de estímulos de comparación cuyo valor estará situado simétricamente en torno al del estándar. Se repetirán varias veces los ensayos y, en cada uno de ellos, el sujeto deberá decir si el estímulo de comparación es mayor o menor que el estándar. Este método es el más utilizado para el cálculo de los umbrales absoluto y diferencial (ver Barbero 1993/1999; 2007 para el cálculo empírico de los umbrales).

La ley de Fechner, a pesar de su indudable importancia, pronto recibe varias críticas debidas, fundamentalmente, a que los estímulos de muy alta o muy baja intensidad no se ajustan bien a ella y, por otra parte, tampoco lo hacen todos los sistemas sensoriales. No obstante, es indudable que introdujo una nueva forma de medición en Psicología, de ahí que se considere que con Fechner comienza la etapa de la Psicología cuantitativa que se ha aplicado a toda clase de problemas psicológicos. Sus trabajos, junto con los que Wundt llevó a cabo en su laboratorio de Psicología fundado 19 años después, marcaron el comienzo de la *Psicología experimental*.

Las controversias que surgen a partir de la Psicofísica desarrollada por Fechner influyeron en otra de las figuras más relevantes en el campo de la Psicofísica: Stevens (1906-1972), quién hace una reformulación de la Psicofísica fechneriana desarrollando los métodos de escalamiento que esta nueva reformulación requería, y una nueva función conocida con el nombre de *Función potencial de Stevens:*

$$R = q \cdot E^n \qquad [1.4]$$

Siendo:

R = respuesta de los sujetos.

E = valor del estímulo.

q = constante que depende de las unidades de medida.

n = exponente de la función que depende del atributo sensorial.

Los trabajos llevados a cabo por Stevens contribuyeron, en gran medida, al resurgimiento de la investigación en el campo de la Psicofísica pues ponían a prueba y comparaban los dos tipos de funciones, la logarítmica y la potencial.

A los métodos desarrollados por Stevens se les conoce también como *métodos directos de escalamiento*, ya que en lugar de originar una escala de sensación (S), como ocurría con los métodos de Fechner, dan lugar a una escala de respuesta (R) cuyos valores son las estimaciones subjetivas que, de forma directa, hacen los sujetos de los estímulos presentados, y cuyas propiedades van a estar determinadas por las instrucciones dadas a los sujetos, a los que se considera capaces de hacer estimaciones a nivel de intervalo y razón.

Stevens trata de medir por un lado el estímulo (E) y por otro la respuesta que de forma directa emiten los sujetos para, posteriormente, encontrar una función que los relacione. Según Stevens esta función es una función potencial. A diferencia de Fechner, no basa sus mediciones en ninguna suposición acerca de las (dap), que Fechner utiliza como unidad de medida de su escala de sensación, ni en el concepto de umbral.

Entre los métodos de escalamiento que utiliza Stevens, merecen citarse: *métodos de emparejamiento de magnitudes* (por modalidad cruzada, estimación de magnitudes y producción de magnitudes), *métodos de emparejamiento de razones* (por modalidad cruzada, estimación de razones y producción de razones), *métodos de emparejamiento de intervalos* (por modalidad cruzada, estimación de intervalos y producción de intervalos) *y las escalas de categorías o de clasificación (Rating scales).*

En los **métodos de emparejamiento de magnitudes por:**

Modalidad cruzada:

Se utilizan dos continuos físicos diferentes y la tarea del sujeto consistirá en emparejar uno con el otro. El experimentador selecciona una serie de estímulos, que varían a lo largo de un continuo determinado cuya magnitud se puede medir mediante los instrumentos adecuados, por ejemplo la intensidad de un sonido, y la tarea de cada sujeto consistirá en ajustar, a cada uno de los estímulos presentados, otro estímulo que varía a lo largo de otro continuo, por ejemplo la separación entre dos rectas verticales y paralelas. Para obtener los valores, tanto de la escala correspondiente al estímulo (escala E) como los correspondientes a la respuesta del sujeto (escala R), se dispone de los instrumentos de medida adecuados.

Estimación de magnitudes:

Se presenta a cada sujeto un estímulo y se le advierte que se fije en él puesto que va a servir de estímulo estándar a partir del cuál habrá de estimar los valores correspondientes al resto de los estímulos que se le van a presentar. En este método, puede ser el experimentador el que asigne un módulo al estímulo estándar y el sujeto irá asignando valores al resto de los estímulos que se le vayan presentando, tomando como referencia el valor asignado al estímulo estándar. Supongamos que se está haciendo un estudio sobre la percepción de los sujetos acerca de la longitud de una serie de líneas y al estímulo que se presenta como estándar se le asigna un módulo de 8 cm. Si el estímulo presentado a continuación le parece al sujeto que es la mitad de largo que el presentado como estándar deberá asignarle un valor de 4 cm si, por el contrario, le parece el doble de largo deberá asignarle un valor de 16 cm.

Producción de magnitudes:

En este método la tarea del sujeto es inversa respecto al método anterior. El experimentador va presentando al sujeto una serie de números, de uno en uno y de forma aleatoria, y la tarea del sujeto consiste en modificar la magnitud de los estímulos en base a los números presentados por el experimentador, de manera que cada número lleve emparejado un estímulo cuya magnitud será, a juicio del sujeto, la representada por el número.

En los **métodos de emparejamiento de razones por:**

Modalidad cruzada:

Se presentan al sujeto dos estímulos que guardan una determinada proporción entre ellos y se le pide que ajuste otros dos, pertenecientes a otro continuo diferente, de manera que guarden entre sí la misma proporción que guardaban los dos primeros. Supongamos que el experimentador le presenta al sujeto dos rectas de la misma longitud y le pide que produzca dos ruidos cuya intensidad mantenga la misma proporción que la que mantenía la longitud de las rectas entre sí; es decir, que sean de la misma intensidad.

Estimación de razones:

Se le presentan al sujeto todos los pares de estímulos, y su tarea consiste en hacer estimaciones de las razones que hay entre las magnitudes de cada par y asignar un número a cada par que represente esa razón. Se trata de asignar razones numéricas a las razones entre las magnitudes de los estímulos. Si mantenemos el ejemplo de la longitud entre dos rectas y en uno de los pares que se le presentan al sujeto la primera línea le parece que es la mitad de larga que la segunda, deberá asignar a ese par el valor numérico de ½, ya que esta razón numérica es la que, a juicio del sujeto, representa la razón entre la longitud de las líneas que forman el par presentado.

Producción de razones:

Se presenta al sujeto un estímulo estándar y junto a él una proporción numérica. La tarea del sujeto consiste en producir otro estímulo que guarde con el estándar una proporción igual a la presentada. Supongamos que al sujeto se le presenta una línea recta de una longitud determinada y se le pide que produzca un estímulo cuya longitud sea la mitad que la del anterior, o bien el doble, etc.

En los **métodos de emparejamiento de intervalos por:**

Modalidad cruzada:

Dados una serie de estímulos que varían a lo largo de dos continuos divididos en intervalos el sujeto habrá de emparejar los intervalos existentes entre los estímulos de un continuo con los intervalos existentes entre los estímulos del otro continuo.

Estimación de intervalos:

Se le presentan al sujeto una serie de estímulos diferentes entre sí y se le pide que, mediante números, haga una estimación de las diferencias estimulares.

Producción de intervalos:

Se le presentan al sujeto dos estímulos y su tarea consiste en encontrar un estímulo intermedio entre los dos presentados (bisección), una serie de estímulos que dividan el intervalo entre los dos estímulos presentados en más de dos intervalos iguales (equisección) o en una serie de intervalos distintos (multisección).

El método de **escalas de categorías o clasificación** es uno de los más utilizados tanto en Psicología como en Sociología. Consiste en asumir que el continuo a lo largo del que se han de situar los estímulos está dividido en una serie de categorías ordenadas cuyos límites serán fijos salvo por error aleatorio. La tarea a realizar será estimar los valores escalares de los límites de las categorías para, una vez hecho esto, asignar los estímulos a cada una de ellas y averiguar sus valores escalares.

A diferencia de los métodos de Fechner que permitían medir la capacidad de detección (mediante umbrales absolutos) y de discriminación (mediante los umbrales diferenciales) de las personas, los métodos desarrollados por Stevens están centrados en medir la capacidad de los sujetos para hacer estimaciones subjetivas acerca de la magnitud de los estímulos y comprobar hasta qué punto los juicios emitidos (estimaciones subjetivas) se ajustan a los datos reales (magnitud real de los estímulos). De esta manera se puede formalizar algo que estamos haciendo constantemente en nuestra vida diaria y que, sin embargo, no siempre somos conscientes de ello. Cuántas veces hemos dicho u oído frases como las siguientes:

Mi casa mide unos 300 metros cuadrados.

Hace aproximadamente dos horas.....

Pues bien, lo que se pretende es comprobar hasta qué punto esas respuestas emitidas por los sujetos se ajustan a la realidad de los datos.

7.2. Métodos de escalamiento psicológico

Los métodos de escalamiento se desarrollaron, en principio, para su utilización en el campo de la percepción pero utilizando características de los estímulos que variaban a lo largo de alguna dimensión física y que, por lo tanto, podían ser medidas con los instrumentos adecuados; se trataba de establecer una relación entre las características físicas de los estímulos y las sensaciones que producían. Ahora bien, ¿qué ocurriría si la característica que se quisiera escalar no variara a lo largo de ningún continuo físico? Si, por ejemplo, se quisiera medir el grado de realismo de una serie de cuadros, las preferencias políticas de una muestra de sujetos, la agresividad, la actitud de los españoles ante la inmigración, etc., difícilmente se podría llevar a cabo el proceso de medición utilizando ninguna escala física, puesto que estas características no varían a lo largo de ningún continuo físico sino a lo largo de un continuo psicológico.

Al surgir este problema comienzan a desarrollarse a finales del siglo XIX, y sobre todo durante el siglo XX, una serie de métodos de escalamiento que aunque basados en las ideas de Fechner tienen unas características propias, como son la no necesidad de recurrir a medidas de tipo físico. A este tipo de métodos de escalamiento se les agrupa bajo la denominación genérica de *métodos de escalamiento psicológico* y a las escalas resultantes *escalas psicológicas o subjetivas*. La figura que más contribuyó al desarrollo de este tipo de escalas fue Thurstone (1887-1955).

Thurstone nació en Chicago de padres suecos, y después de obtener el título de ingeniero civil en la universidad de Cornell se dedicó a la electrotecnia. Antes de graduarse había patentado un modelo de proyector de películas que atrajo la atención de Edison quien le ofreció un puesto en su laboratorio de East Orange (Nueva Jersey), pero Thurstone pronto abandona el laboratorio para dedicarse a enseñar geometría y dibujo. En 1914 se interesa por el estudio experimental del aprendizaje y se matricula en Psicología en la Universidad de Chicago siendo ayudante de Bingham en el Departamento de Psicología Aplicada del Instituto Carnegie de Tecnología, del que llegó a ser Director después de doctorarse.

A finales de la década de 1920, L. L. Thurstone estaba en la universidad de Chicago como profesor de Psicofísica, pero convencido de que las sensaciones que medía con tanta precisión no merecían su tiempo y esfuerzo, y aburrido de todo aquello que enseñaba, decidió probar fortuna y medir, aplicando los mismos métodos, algo que para él tuviera importancia.

En lugar de presentar a los sujetos dos objetos y preguntarles, por ejemplo, ¿cuál de estos dos objetos es más largo?, se les podrían presentar dos cuadros y preguntarles ¿cuál de estos dos cua-

dros tiene más realismo?, o bien, tal y como hizo, presentarles una serie de delitos y pedirles que los diferenciaran en función de su gravedad. Si este tipo de problemas se pudiera tratar mediante algún tipo de lógica psicofísica, se abriría la posibilidad de una descripción objetiva de mayor significación psicológica que el umbral sensorial (Miller, 1982).

Thurstone trató de elaborar un modelo a partir del cual pudiera elaborar una escala sobre un continuo psicológico y situar en ella los estímulos (también psicológicos) sin necesidad de recurrir a ninguna operación en un continuo físico. El modelo que desarrolló está basado, por una parte, en la variabilidad perceptual de los sujetos, e incluso en la de un mismo sujeto cuando se le presentan los mismos estímulos en distintas ocasiones y, por otra, en la limitación que tienen los sujetos para percibir las diferencias de magnitud entre dos estímulos cuando estas son muy pequeñas. En la medida en que la diferencia entre la magnitud de los estímulos sea mayor, será más fácil que los sujetos puedan diferenciarlos y ordenarlos respecto a la característica o atributo que se esté evaluando; mientras que, en la medida en que los estímulos sean más parecidos, los sujetos encontrarán mayores dificultades para realizar su tarea. Thurstone (1927) publicó varios trabajos que trataban diversos problemas acerca de la medición subjetiva o psicológica, y desarrolló un modelo matemático, relacionado con la Psicofísica clásica de Fechner, cuyas ecuaciones se conocen con el nombre de *Ley del Juicio Comparativo*. Posteriormente desarrollaría otro modelo cuyas ecuaciones se conocen como *Ley del Juicio Categórico*.

Estos dos modelos llevan asociados una serie de métodos experimentales para la obtención empírica de los datos; la Ley del Juicio Comparativo utiliza el *Método de las Comparaciones Binarias*. La Ley del Juicio Categórico: el *Método de los Intervalos Sucesivos*, el *Método de los Intervalos Aparentemente Iguales* y el *Método de Ordenación de Rangos* fundamentalmente.

Aunque Thurstone atribuyó a otros la fundación de la *Psychological Society* y de su revista *Psychometrika*, se hallaba muy unido al grupo de personas que en 1936 crearon la sociedad y la revista para fomentar el desarrollo de la Psicología como ciencia racional cuantitativa.

A partir de los trabajos de Thurstone se fueron desarrollando nuevas formas de escalamiento psicológico, así:

Guttman en los años 40 del siglo XX, desarrolló un nuevo modelo para el escalamiento conjunto de sujetos y estímulos (escalamiento de respuestas). Para la obtención de la escala utilizó un método conocido como *método del escalograma*. La escala resultante se denomina *escala de entrelazamiento* puesto que tanto los sujetos como los estímulos se sitúan a lo largo del mismo continuo psicológico de forma entrelazada.

Coombs (1950), desarrolló una teoría conocida como *Teoría del despliegue* y propuso un modelo que, al igual que el de Guttman, permite escalar sujetos y estímulos conjuntamente.

Todos estos modelos de escalamiento y sus métodos asociados se desarrollaron, en principio, para la construcción de escalas unidimensionales; es decir, de escalas que permitieran ordenar un conjunto de estímulos, o estímulos y sujetos en los dos últimos casos, respecto a un único atributo o característica mediante la asignación de un único valor escalar que represente la posición del objeto escalado.

En los años 60 se desarrollaron los métodos de *escalamiento multidimensional*, aunque, en realidad, los orígenes de estos métodos haya que situarlos en 1938 en los tratados de Young y Hoseholder y los de Richardson. A diferencia de los métodos de escalamiento unidimensionales, estos métodos permiten la ordenación de los objetos a escalar atendiendo simultáneamente a más de un atributo o característica y, por lo tanto, asumen la existencia de más de una dimensión subyacente al conjunto de observaciones. En este caso, en lugar de asignar un único valor en la escala a cada uno de los objetos, se les asignará un valor en cada una de las dimensiones analizadas.

Hay tres obras clásicas sobre escalamiento, tanto psicofísico como psicológico. La obra de Guilford *Psychometric Methods* publicada en 1954; la de Torgerson *Theory and Methods of Scaling* de 1958 y el libro de Edwards *Techniques of Attitude Scale Construction* de 1957.

En castellano se puede consultar el libro de Barbero (1993/1999; 2007) y el de Blanco (1996).

3. ORÍGENES Y DESARROLLO DE LOS TESTS

Como se comentó al inicio del tema, el segundo problema que motivó la introducción del proceso de medición en Psicología fue el intento de apreciar de forma sistemática las diferencias individuales; es decir, el intento de escalar a los sujetos. Este intentó llevó al desarrollo del *Método de los Tests* y de las distintas *Teorías de los Tests*.

Podríamos remontarnos bastantes años antes de Cristo y ya se apreciaría el interés existente por analizar de alguna manera las diferencias individuales. En China se utilizaban tests para seleccionar a las personas que ocuparían puestos en el gobierno. Autores clásicos como Platón e Hipócrates propusieron también algunas formas para conseguir analizar las diferencias individuales y, de un modo más concreto, en el siglo XVI lo hace el español Huarte de San Juan.

En 1796 Kinneybrook, uno de los investigadores que trabajaba en el Observatorio Astronómico de Greenwich, fue expulsado porque discrepó del resto de sus compañeros en la estimación que hizo del tiempo que tardaba una estrella en cruzar un determinado espacio, y se consideró que se había equivocado en los cálculos. Hubo que esperar hasta el año 1822 a que los astrónomos comprendieran que las personas tienen diferentes tiempos de reacción y que, por lo tanto, a la hora de interpretar los datos observados era necesario tener esto en cuenta (Freeman, 1926), y hubo que esperar hasta la mitad del siglo XIX para que se desarrollaran en Europa y en Estados Unidos procedimientos formalizados para la evaluación de tales diferencias: *los tests*.

Antes de continuar con la reseña histórica del origen y desarrollo de los tests creemos necesario aclarar el significado psicométrico del término. Son muchas las definiciones que han ido apareciendo a lo largo de los años, pero todas ellas hacen referencia a que:

> *Un test es un instrumento de medición diseñado especialmente para estudiar de un modo objetivo y sistemático el nivel de los sujetos respecto a algún atributo, característica o dominio de conocimientos y, a partir de las puntuaciones que obtengan los sujetos en el test, poder analizar las diferencias existentes entre ellos.*

Hay tres factores que se pueden considerar decisivos en el desarrollo de los tests: la apertura del laboratorio antropométrico de Galton en Londres, el desarrollo de la correlación de Pearson y la interpretación que Spearman hace de ella, considerando que la correlación entre dos variables indica que ambas tienen un factor común.

8.1. La importancia de los trabajos de Galton

Francis Galton (1822-1911), era nieto de Erasmus Darwin y medio primo de Charles Darwin, por lo que es fácil comprender la influencia que tuvo la teoría de la evolución en sus trabajos.

Las ideas evolucionistas plasmadas en la obra de Charles Darwin titulada *The Origin of Species by Means of Natural Selection* y publicada en 1859, tuvieron una enorme acogida por parte de Francis Galton pero preocuparon enormemente a Wundt ya que el evolucionismo defendía una filosofía completamente diferente a la suya. Mientras que Wundt trataba de encontrar las leyes generales que dirigen o gobiernan la *mente* humana, los evolucionistas trataban de clasificar y catalogar las diferentes formas de *mentes* que podían existir. Esto fue el desencadenante que propició la separación entre la Psicología Experimental y la Psicología diferencial.

Galton todavía creía que la medición de las características mentales estaba estrechamente relacionada con la de las características físicas, y pensaba que para poder estudiar las dimensiones de la mente del hombre debía obtener la misma clase de datos antropológicos que estaba reuniendo al tratar de estudiar su anatomía. Comprendió que para obtener una descripción completa del hombre, las medidas antropométricas de estatura, peso, color de la piel, etc., deberían completarse con medidas psicométricas de los sentidos, la memoria, etc. Pero, como señala Miller (1982) hasta el día de su muerte fue reacio a admitir que *el tamaño del cráneo de un hombre no tenía valor como medida de su inteligencia.*

En 1884 montó su laboratorio antropométrico en la Exposición Internacional de la Salud celebrada en el Museo de South Kensington en Londres. Allí, por el precio de tres peniques, se tomaban a los visitantes medidas en una variedad de pruebas físicas y sensoriales pues, entre otras cosas, consideraba que la inteligencia podía medirse a través de la actividad sensomotora y, una vez recogidos los datos, comenzó elaborando distribuciones de frecuencias y comprobando que su forma era aproximadamente la misma para las variables psicológicas que para las anatómicas, la distribución normal. Ahora bien, a cada persona se le habían tomado datos de distintas variables y, ante esa cantidad de datos, pronto se plantea el problema de cómo encontrar relaciones entre medidas diferentes. ¿Cuál era la relación entre ellas? Por ejemplo la gente de estatura alta tendía a pesar más que la de estatura baja. ¿Cómo podía medirse esta tendencia? Este problema, que podía tener cierta lógica, se complicaba cuando se trataba de hacer estudios sobre la herencia. ¿Cómo se podían relacionar las mediciones hechas a los padres con las de los hijos? Para solucionar este problema Galton recurrió a la correlación como medida de asociación. Así, Galton fue el primero en aplicar el concepto estadístico de distribución normal, de media, de mediana, varianza y correlación a datos psicológicos. Esta tarea fue continuada por Karl Pearson (Miller, 1982).

Para poder llevar a cabo sus mediciones, Galton construyó y utilizó una serie de instrumentos que pueden ser considerados los primeros *Tests psicométricos*, pero éstos se caracterizaban por un fuerte carácter antropométrico, sensorial y motor. No obstante, a partir de Galton empieza el gran desarrollo del método de los tests.

8.2. Los primeros tests mentales

James McKeen Cattell en 1888, estuvo en la Universidad de Cambridge y allí se puso en contacto con Galton, estableciéndose entre ellos un nexo de unión dado su común interés por investigar las diferencias individuales; no obstante, este trabajo común duró poco tiempo ya que Cattell volvió a Estados Unidos como profesor de Psicología en la Universidad de Pennsylvania. En 1891 se trasladó a Columbia donde fundó un laboratorio de Psicología en el que trabajó durante 26 años, hasta que fue dimitido por sus posiciones pacifistas cuando Estados Unidos entró en la Primera Guerra Mundial en 1917.

La asociación entre Cattell y Wundt en Alemania y de Cattell y Galton en Inglaterra fue, por una parte, el lazo de unión entre los laboratorios psicológicos alemanes e ingleses y, por otra, un hecho que contribuyó al intercambio de ideas entre los investigadores europeos y americanos (French y Hale, 1990).

Cattell utilizó por primera vez el término *Test Mental* en su artículo *Mental Test and Measurement* publicado en 1890 en la revista *Mind*. Pero los tests a los que hace referencia Cattell, al igual que los de Galton, tenían un carácter sensorial y motor fundamentalmente, ya que incluían medidas de energía muscular, velocidad de movimientos, sensibilidad al dolor, etc. El análisis de los datos puso de manifiesto que la correlación entre este tipo de pruebas y el nivel intelectual de los sujetos era nula (Wissler, 1901). Para Cattell los tests constituían *un sistema uniforme que permite comparar y combinar, en lugares y momentos diferentes, la medida de las funciones mentales* (Cattell, 1890;

pág. 374), y compartía con Galton la creencia de que se podía medir el funcionamiento intelectual de las personas mediante tests de discriminación sensorial y midiendo el tiempo de reacción.

Aunque durante las dos últimas décadas del siglo XIX hay una enorme expansión de los estudios acerca de las diferencias individuales en Inglaterra y en Estados Unidos, los tests seguían siendo, fundamentalmente, de tipo sensorial y motor. Fue necesario llegar a finales de siglo para aceptar que estos tests no medían la inteligencia. Las puntuaciones obtenidas por los sujetos en estos tests no guardaban relación con las medidas de rendimiento escolar, lo que venía a refutar la hipótesis de la relación entre la capacidad sensorial y la inteligencia.

8.3. Desarrollo de los primeros tests de inteligencia

Alfred Binet (1857-1911), fue el primero en darse cuenta de que las sensaciones no jugaban un papel demasiado importante en la Psicología diferencial y que era necesario centrarse en el estudio de los procesos mentales superiores. Otro de los aciertos de Binet fue el darse cuenta también de la importancia que tenía la edad de los sujetos como variable interviniente. En 1889 fundó con Beaunis el primer laboratorio de Psicología de la Sorbona del que llegó a ser director en 1894.

Binet trabajó con Simon y realizaron una serie de investigaciones cuyo resultado fue la publicación en 1905 de lo que puede ser considerada la primera escala de inteligencia. A diferencia de los tests de Galton, este test ya no se ocupaba de las funciones motoras o sensoriales sino de la capacidad de comprensión y razonamiento de los niños. Esta escala, conocida como la *Escala de Binet-Simon* constaba de 30 elementos de dificultad creciente, y aunque se incluían algunos elementos de tipo sensorial la mayoría eran verbales. Esta escala fue revisada en 1908, traducida a numerosos idiomas y adaptada varias veces. La adaptación más conocida del test de Binet fue la que realizó en 1916 el psicólogo americano Terman en la Universidad de Stanford, de ahí el nombre de *Test Stanford-Binet de Terman*. Este test fue traducido y adaptado en España por Germain y Rodrigo (1928).

Aunque la finalidad inicial de la escala era detectar a los sujetos que presentaran algún retraso intelectual, posteriormente se utilizó para el estudio de las diferencias en otros niveles. Para poder interpretar las puntuaciones obtenidas, se desarrolló el concepto de *Edad Mental* que equivale a la edad cronológica de los niños intelectualmente normales cuya media en la escala es igual a la puntuación obtenida por el niño examinado. Es decir, se aplica la escala a una muestra representativa de niños intelectualmente normales divididos en distintos niveles de edad y se calcula la puntuación media de los niños en cada uno de los niveles. Esa sería la norma establecida con la cual comparar los resultados obtenidos. Así, supongamos que se aplica la escala a un niño de 12 años y obtiene una puntuación que coincide con la puntuación media obtenida por los niños de 9 años; entonces, diremos que la edad cronológica del niño examinado es de 12 años pero su edad mental es de 9 años.

Terman (1877-1956), consideraba que la inteligencia se podía definir como la habilidad para llevar a cabo razonamientos abstractos y, a pesar de que sigue la táctica de Binet para la elaboración de sus tests, utiliza como medida de la inteligencia el concepto de *Cociente Intelectual* (CI) acuñado por Stern en 1912. El cociente intelectual (no coeficiente intelectual como se escucha muchas veces) equivale al cociente entre la edad mental y la edad cronológica, generalmente multiplicado por cien para evitar los decimales. Como señala Santisteban (1990) al definir el cociente intelectual se establece la primera escala de puntuaciones referida a las aptitudes mentales. En 1937 junto a Merrill llevaron a cabo otra revisión del test de Binet y obtuvieron dos formas paralelas.

8.4. Los tests colectivos

Todos los tests desarrollados hasta el momento eran de aplicación individual, lo que representaba un gran inconveniente por la cantidad de tiempo que requería su aplicación; pero la entrada de EE.UU. en la Primera Guerra Mundial en 1917, y la necesidad de seleccionar y clasificar el contingente de personas disponibles en función de sus capacidades, hacía imposible la aplicación de este tipo de pruebas. Entonces, se nombró una comisión especial de la Asociación Norteamericana de Psicología, dirigida por R.M. Yerkes, con el encargo de investigar nuevos procedimientos que posibilitaran la administración de tests en el ejército. Para ello, se pusieron en contacto con *Arthur Otis* que llevaba tiempo trabajando en un test colectivo de inteligencia y, a partir de ese material, se diseñaron los famosos tests conocidos como *Tests Alpha y Beta* que fueron publicados en 1918. El primero, el Test Alpha, diseñado para la población general y el segundo, el Test Beta, para los reclutas analfabetos o aquellos que no dominaban el inglés. Ambos tests siguen usándose hoy día tras numerosas revisiones. Este fue el comienzo de los tests colectivos (Otis, 1939).

Una vez finalizada la guerra el uso de los tests se extendió a la industria y al resto de las instituciones. En 1922 Catell fundó la primera empresa dedicada a la publicación masiva de tests, y a partir de ese año empezaron a utilizarse normalmente como instrumento de selección en la administración americana. Su sistema de valoración era sencillo; en general, se asumía como puntuación de los sujetos el número de aciertos y para su interpretación se recurría, bien a normas cronológicas (edad mental o cociente intelectual), bien a normas estadísticas (centiles o puntuaciones típicas).

En 1938 apareció el *Test de Weschler-Bellevue* para la medición de la inteligencia en los adultos y en 1949 apareció una versión para niños conocida, de forma abreviada, como *WISC*. En 1955 apareció una revisión de la escala anterior denominada *Weschler Adult Intelligence Scale (WAIS)* que, a su vez, fue revisada en 1981, y en 1967 apareció el *Weschler Preschool and Primary Scale of Intelligence*. La aparición de estas escalas supuso un gran avance en el desarrollo de los tests, entre otras cosas, porque se había constatado la falta de efectividad que tenía el test de Stanford-Binet aplicado a adultos.

A pesar del gran desarrollo del método de los tests, quedaban una serie de preguntas sin respuesta: ¿qué es lo que miden realmente los tests?, ¿existen realmente los rasgos que dicen medir los tests? En un intento de dar respuesta a estos interrogantes se desarrollaron el coeficiente de correlación de Pearson y una serie de técnicas correlacionales conocidas bajo el nombre genérico de *Análisis Factorial*. Los primeros esbozos de estas técnicas se deben a Spearman y hay que resaltar la enorme importancia que tuvieron en el desarrollo del método de los tests, ya que sin estas técnicas los tests mentales hubieran tardado más en perder su carácter básicamente sensorial.

El objetivo común de las técnicas reunidas bajo el término *Análisis Factorial* es representar un conjunto de variables observables (escalas, ítems, etc.) mediante un número más pequeño de variables no observables sino latentes llamadas factores de las que las primeras son indicadores.

La influencia del Análisis Factorial respecto a los tests de inteligencia fue doble. Por una parte dio una fundamentación teórica a su utilización, mostrando que en la mayor parte de las funciones cognoscitivas interviene un factor general «g» común a todas ellas. Por otra, los resultados del análisis factorial subrayan la importancia de estudiar otras aptitudes más específicas en el campo de la inteligencia. Surgen así los primeros tests destinados a la medida de las aptitudes y del rendimiento.

8.5. Las baterías de aptitud múltiple

Los tests de inteligencia utilizados hasta el momento proporcionaban información acerca del nivel intelectual global de los sujetos y eso no bastaba, era preciso trazar su perfil mental en función de las puntuaciones específicas obtenidas en diversos factores cognoscitivos. Así, a partir de los años treinta, y sobre todo a partir de los cuarenta del siglo XX, cuando ya la técnica del análisis factorial empezaba a dar sus frutos, comenzaron a desarrollarse las baterías de tests (*Baterías de Aptitud Múltiple*), cuya finalidad era procurarnos una medida de la posición de cada sujeto en un cierto número de rasgos. En lugar de una puntuación total, o CI, se obtiene una puntuación por separado para cada rasgo, dando, como señala Muñiz (1992), una mayor importancia a un factor general de inteligencia que articularía jerárquicamente otros factores de grupo (escuela inglesa), o reclamando un plano de igualdad para factores múltiples (escuela americana).

Entre las primeras baterías merece destacar la *Batería de Aptitudes Mentales Primarias* de Thurstone (Thurstone, 1938; Thurstone y Thurstone, 1941) que incluía pruebas para la medida de lo que entonces se consideraban los componentes fundamentales del comportamiento inteligente: comprensión verbal, fluidez verbal, aptitud numérica, aptitud espacial, memoria, rapidez perceptiva y razonamiento general. Hoy día existen numerosas baterías (PMA, DAT, etc.) de uso habitual.

8.6. Los tests de personalidad

Aunque la medida de las diferencias individuales había estado centrada fundamentalmente en el campo de las facultades mentales; se habían hecho algunos intentos de extender el uso de los tests a otros campos de la Psicología: la personalidad, los intereses, las actitudes, etc. Se puede considerar a Kraepelin como un precursor de los *Tests de Personalidad,* pues ya en 1894 utilizó el *Test de Asociación Libre* con pacientes normales para estudiar los efectos psicológicos de la fatiga, el hambre y las drogas, y llegó a la conclusión de que todos esos agentes aumentan la frecuencia relativa de las asociaciones superficiales (Anastasi, 1980).

El prototipo de cuestionario de personalidad con carácter psicométrico, es la *Hoja de Datos Personales* de Woodworth (1917), utilizada en la Primera Guerra Mundial y cuya finalidad era la detección de aquellos sujetos que por padecer trastornos neuróticos graves no eran aptos para el servicio militar. Posteriormente se prepararon distintas formas de este cuestionario e incluso se preparó una forma para niños.

De los primeros tests objetivos de personalidad hay tres que han sobrevivido en la actualidad: una serie desarrollada por Guilford y sus colaboradores (Guilford y Zimmerman, 1949); una serie similar desarrollada por Cattell y sus colaboradores (Cattell, Eber y Tatsuoka, 1970) y el *Inventario Multifásico de Personalidad de Minnesota* (MMPI) de Hathaway y McKinley (1942). En general los tests de Guilford y de Cattell se usan en «poblaciones normales» mientras que el MMPI se usa en «poblaciones clínicas». Además, los tests de Guilford y de Cattell están basados en el análisis factorial y están orientados hacia los rasgos mientras que el MMPI, en su forma estándar, no utilizaba escalas derivadas del análisis factorial y está orientado hacia una clasificación psiquiátrica.

Durante la primera mitad del siglo XX se fueron perfilando otro tipo de tests, los *Tests Proyectivos;* así, en 1921, el psiquiatra suizo Rorschach diseñó el primer test que pretendía dar una visión global y comprensiva de la personalidad, el *Test de Psicodiagnóstico*, se trata del famoso test proyectivo también conocido como *el Test de las manchas de tinta*. A partir de este test se desarrollaron otros muchos que pronto se vieron asociados con la Psicología clínica y, en especial, con el Psicoanálisis. En 1926 aparece el test de *Dibujo de la Figura Humana* de Machover, en 1938 el *Test de Apercepción Temática* (TAT), etc.

No obstante, como señalan Goldstein y Hersen (1984), hay varias razones que han contribuido a que disminuya la utilización de estas técnicas: *a)* el aumento de la sofisticación científica ha creado una atmósfera de escepticismo en torno a estos instrumentos, *b)* el desarrollo de procedimientos alternativos, por ejemplo el MMPI y otros tests objetivos, convenció a algunos psicólogos clínicos de que a partir de estos procedimientos se podía obtener la misma información y de una manera menos costosa y *c)* las técnicas proyectivas, en general, estaban asociadas en cierta medida con la teoría psicoanalítica. A pesar de estas razones científicas, prácticas y filosóficas el test de Rorschach sigue mostrando su utilidad.

8.7. La medición de intereses y actitudes

Entre los instrumentos clásicos para la medida de los intereses merecen citarse el *Cuestionario de Intereses Vocacionales de Strong (SVIB)* desarrollado por E.K. Strong (1927) y la *Escala de Preferencias de Kuder* introducida por Kuder en 1934. La más reciente revisión de esta escala es la efectuada en 1985.

En cuanto a la medida de las actitudes, creencias y opiniones, su desarrollo efectivo tuvo lugar a partir de 1928, fecha en la que Thurstone desarrolló sus dos modelos, la *Ley del Juicio Comparativo* y, sobre todo, la *Ley del Juicio Categórico*, junto con los procedimientos experimentales para la recogida de datos que, basados en los métodos psicofísicos desarrollados por Fechner iban a permitir el escalamiento de los ítems (estímulos). La técnica utilizada por Thurstone para la medida de las actitudes (Thurstone y Chave, 1929), tal y como se verá más adelante, permite la ordenación de los ítems (estímulos) a lo largo de una escala continua en función de los valores escalares que les hayan asignado el grupo de sujetos a los que se les han presentado para su evaluación (prueba de jueces); una vez escalados los ítems se podrá utilizar la escala para averiguar el grado de actitud de los sujetos en la variable medida; es decir, para escalar a los sujetos.

Algunos años más tarde Likert (1932) desarrolló una nueva técnica que vino a paliar alguno de los inconvenientes de la de Thurstone, entre otros la prueba de jueces. Las escalas elaboradas siguiendo la *Técnica de Likert*, están basadas en los mismos principios que la mayoría de los tests de aptitudes. Algunos investigadores, entre los que podemos citar a Edwards y Kenney (1946), encontraron que este tipo de escalas era más fácil de elaborar que las de Thurstone y diagnosticaban mejor.

Con el fin de medir el significado connotativo, también llamado significado afectivo o subjetivo, que determinados estímulos tienen para los sujetos, Osgood (1952) revisó todas las teorías del significado y así pudo encontrar el marco teórico que le permitió desarrollar una escala de clasificación para medirlo: *El Diferencial Semántico*.

> **Nota:** Dada la importancia de las contribuciones de Thurstone no vamos a extendernos aquí en la explicación de sus modelos y los métodos que llevan asociados. En el Tema 3 se hará una exposición algo más detallada de los mismos, junto a otras técnicas como las de Likert, Guttman y Osgood, por su importancia para la medida de las actitudes.

8.8. Institucionalización del uso de los tests

Durante los años cuarenta y cincuenta del siglo xx se produjo un aumento considerable en el uso de los tests pero también un cierto abuso, lo que condujo a numerosas críticas sociales e incluso a la prohibición de su uso en determinados estados. Yela (1977) cuenta cómo en Estados Unidos se aplicaron durante 1946 unos sesenta millones de tests a cerca de veinte millones de personas. A las críticas sociales provocadas por los excesos en el uso de los tests se unieron ciertas críticas aparecidas en la literatura científica acerca de la adecuación de los mismos para ciertos usos y colectivos, dado que muchos de ellos presentaban determinados sesgos (Jensen, 1980; Levine y Rubin, 1979; Lumsden, 1976; Pawlik ,1976). No obstante, como señala Santisteban (1990) las limitaciones de los tests no justificaron del todo esta fuerte corriente crítica, puesto que la teoría y la práctica demostraron que:

> *... las deficiencias que a priori parecen ser propias de los instrumentos que tratan de medir capacidades humanas complejas a través de simples manifestaciones observables, son generalmente conocidas y controlables y que, por lo tanto, basta considerarlas para hacer un uso correcto del test. Asimismo, habrá que tenerlas en cuenta al hacer la interpretación de las puntuaciones, puesto que es obvio, en cualquier proceso de medición, el que la interpretación de la medida se ajuste también a las características del instrumento* (pág.3).

De acuerdo con Meliá (1990) la cuestión puede consistir en no confundir la teoría psicométrica con los tests derivados de ella, ni los tests con las teorías psicológicas ni con los usos inadecuados que se hagan de los mismos.

El uso masivo de los tests fue extendiendo su campo de aplicación a distintos sectores tanto públicos como privados, en la selección de puestos en el gobierno y en las escuelas, para ayudar a niños con problemas escolares, en orientación, clasificación, etc., lo que ocasionó la aparición de instituciones tanto públicas como privadas dedicadas a ello. El aumento de estudiantes que querían acceder a la escuela secundaria hizo necesario un plan de selección, así en 1899 se formó el *College Entrance Examination Board (CEEB)* quienes desarrollaron una serie de tests para realizar dicha selección. Los primeros exámenes del CEEB estaban basados en contenidos curriculares, pero hacia el año 1928-29 se desarrolló el *Scholastic Aptitude Test (SAT)* que intentaba medir la *capacidad para el aprendizaje* más que lo que ya habían aprendido los estudiantes.

En 1947, el CEEB, el *American Council on Education* y otras instituciones se unieron para fundar el *Educational Testing Service* (ETS) con el fin de potenciar la investigación en el campo educativo. El ETS jugó un importante papel en el desarrollo histórico de la Psicometría como soporte y núcleo de formación de una gran parte de las principales figuras extranjeras que hoy día existen en este campo y de algunas españolas que han ido allí y siguen yendo a formarse.

En 1951 fueron nombrados el comité de la *American Psychological Association* (*APA*), el de la *American Educational Research Association (AERA),* y el *National Council on Measurements Used in Education (NCMUE* más tarde *NCME)* para preparar una serie de recomendaciones técnicas para el uso de los tests, tal y como se verá más adelante. Inicialmente, intentaron preparar cuatro

manuales independientes sobre aptitudes, rendimiento, intereses y personalidad basados en los trabajos de los diferentes comités, pero debido a la similitud de las recomendaciones se decidió que la APA publicara un volumen sobre aptitudes, intereses y personalidad y que la AERA y el NCME publicaran uno sobre rendimiento a través del *National Education Association*, (APA, comité sobre tests psicológicos, 1955). Publicación que constituye un punto de referencia común.

En la publicación llevada a cabo en 1999, la *American Psychological Association* (*APA*), la *American Educational Research Association (AERA),* y el *National Council on Measurements in Education (NCME),* publicaron los *Standards for Educational and Psychological Testing,* con el fin de proporcionar a los profesionales de la Psicometría criterios para la evaluación de los tests. La última edición es la de 2014.

En España, y con el fin de proporcionar una serie de directrices que ayuden a mejorar el uso de los tests por parte de los profesionales, se creó en 1995 la *Comisión de Tests* por el Colegio Oficial de Psicólogos, comisión que funciona de forma coordinada con otras comisiones internacionales creadas con la misma finalidad entre las que merecen destacar la *Task Force* sobre tests de la *Federación Europea de Asociaciones profesionales de Psicólogos (*EFPPA), o la *International Tests Commission* (ITC).

8.9. Los tests referidos al criterio frente a los referidos a normas

A finales de los años sesenta y durante la década de los setenta del siglo xx, se había hecho en los EE.UU una fuerte inversión económica en el campo educativo y, lógicamente, se deseaba conocer los resultados de la misma para poder averiguar si se había conseguido que los alumnos alcanzaran unos objetivos mínimos (estándares educativos) y, a la vez, evaluar su nivel de competencia y sus habilidades básicas.

Los tests que se utilizaban habitualmente no se adecuaban bien a los nuevos objetivos planteados, ya que se habían desarrollado para evaluar el grado de aptitud o el nivel de los sujetos en un determinado rasgo, pero interpretando los resultados en relación a los que se habían obtenido en una muestra representativa que había servido para establecer una serie de normas (grupo normativo), sin analizar si los sujetos habían alcanzado o no los objetivos mínimos a nivel educativo: *Tests Referidos a las Normas (TRN)*. Por el contrario, el interés del momento estaba centrado, no en la evaluación del nivel de rasgo o aptitud de los sujetos, sino en la evaluación del rendimiento y en el diagnóstico de las necesidades que pudieran tener los alumnos de cara a una posible intervención. Se trataba de averiguar hasta qué punto los alumnos dominaban el contenido de determinadas materias o eran capaces de resolver correctamente problemas habituales de su vida real, de ahí que se desarrollara otro tipo de tests, *los Tests Referidos al Criterio (TRC)* que iban a tratar de dar solución al problema.

Estos tests proporcionaron las bases para interpretar las puntuaciones de los sujetos en relación a un *dominio* bien definido, más que en relación a un grupo normativo que era como se venía haciendo, y permitieron a los psicólogos y educadores la evaluación de los estudiantes en función de su conocimiento o no de una materia determinada, en lugar de hacerlo en relación a otros estudiantes (Berk, 1984; Hambleton, 1985).

Hay un acuerdo generalizado en considerar que Glaser (1963) fue el primero que distinguió entre las dos aproximaciones a la medición del rendimiento: la referida a la norma y la referida al criterio y estableció que

> *...los TRC son aquellos en que los resultados dependen del estatus absoluto de calidad del estudiante, frente a los TRN que dependen del estatus del grupo. (pág. 519).*

La definición más aceptada en la actualidad es la de Popham (1978) para quien

> *...un test referido al criterio se utiliza para evaluar el estatus absoluto del sujeto con respecto a algún dominio de conductas bien definido* (pág. 93).

Los trabajos de Glaser suscitaron un enorme interés, y a partir de ese momento aparecieron muchas publicaciones sobre el tema.

Las referencias a trabajos realizados sobre este tipo de tests es continua, sobre todo en algunas revistas como *Journal of Educational Measurement, Review of Educational Research, Applied Psychological Measurement, American Educational Research y Journal of Educational Statistics.*

Las diferencias entre los TRN y los TRC no siempre son aparentes a simple vista ya que ambos tipos de tests están formados por ítems de formatos similares, requieren el mismo tipo de operaciones cognitivas de los sujetos, etc. Sin embargo, como recoge Martínez-Arias (1995, pág. 657) hay grandes diferencias tanto en su construcción como en la interpretación de las puntuaciones obtenidas por los sujetos. Estas diferencias hacen referencia a cinco aspectos fundamentalmente:

— *La finalidad de la evaluación*:

Mientras que en los TRN el objetivo es poner de manifiesto las diferencias individuales en la conducta o rasgo que mide el test, en los TRC el objetivo es estimar el rendimiento o conducta del sujeto en los objetivos que mide el test.

— *La construcción del test y la especificación de los contenidos*:

En los TRN a la hora de construir los elementos que formarán el test se suele recurrir a las teorías existentes respecto al rasgo o constructo que se quiere medir, pero no suelen tener una delimitación clara del dominio de contenidos que se van a evaluar. Por el contrario, en los TRC lo primero que hay que hacer es especificar de una manera clara el dominio de contenidos o conductas que se quieren evaluar y el uso que se pretende hacer del test.

— *La forma de seleccionar los ítems:*

En los TRN los ítems deben poner de relieve las diferencias individuales maximizando la varianza del test, y seleccionando ítems de dificultad media y alto poder discriminativo. En los TRC los ítems se seleccionan en función de los objetivos y del uso que se vaya a hacer del test.

— *El significado de las puntuaciones*:

En los TRN la puntuación obtenida por los sujetos se considera un indicador de su puntuación verdadera en un rasgo latente. En los TRC la puntuación es un estimador de la conducta o rendimiento del sujeto en el dominio.

— *La interpretación de las puntuaciones*:

Mientras que en los TRN la puntuación tiene significado únicamente con relación a los resultados del grupo normativo, en los TRC la puntuación tiene significado en términos absolutos.

Estas diferencias implican que a nivel psicométrico se mantenga esta diferenciación entre los dos tipos de tests.

8.10. Los tests adaptativos informatizados (TAI's)

Los avances en el campo de la informática y el hecho de que el ordenador haya bajado su coste de manera ostensible han permitido desarrollar nuevos métodos de enseñanza-aprendizaje y de evaluación del progreso. Asimismo, se han modificado los procedimientos de administración de tests; cada vez son más populares las aplicaciones a través del ordenador. Un ejemplo de este tipo de aplicaciones lo constituyen los *Tests Adaptativos Informatizados (TAI's),* denominados así porque la selección de los ítems se va haciendo a lo largo del proceso de administración del test, de manera que los ítems se puedan *adaptar* al nivel de cada sujeto y no le resulten ni demasiado fáciles ni demasiado difíciles. Cuando se administran estos tests de forma computerizada, se utiliza el ordenador como medio para hacer la selección de los ítems (Larkin y Weiss, 1974; Olea y col., 1999; Weiss, 1983).

Teniendo en cuenta el interés de la Psicología cognitiva por el estudio de la forma en que las personas procesan y representan mentalmente la información recibida, es fácil darse cuenta que muchos de los procesos medidos tradicionalmente por medio de los tests psicométricos (percepción, memoria, razonamiento, toma de decisión, etc.) entran dentro de su dominio. Este hecho ha sido la causa de que muchos psicómetras se hayan planteado estudiar nuevas formas de construcción de tests de manera que se tengan en cuenta los avances habidos en este terreno.

El lector interesado en conocer las implicaciones que estos desarrollos han tenido para la construcción de tests pueden consultar los trabajos de Embretson (1985, 1993); Messick, (1989); Mislevy (1993); Pellegrino (1988); Prieto y Delgado (1996); Sternberg (1981, 1988, 1991). Una revisión de los trabajos sobre Inteligencia y Procesamiento de la Información puede encontrarse en Martínez-Arias (1991).

Una exposición más detallada del origen y desarrollo de los tests puede verse en Anastasi (1988), Du Bois (1970), Muñiz, (1998), Nunnally (1970), y Thompson y Sharp (1988) entre otros. Meliá (1990) realiza una aproximación bibliométrica a la Psicometría bastante exhaustiva y una información detallada sobre tests concretos se puede encontrar, entre otros sitios, en las ediciones sucesivas de los *Mental Measurement Yearbooks* editados por Buros que incluyen varios volúmenes con miles de tests incluidos y en las páginas Web de distintas editoriales. Estas publicaciones ponen de manifiesto la gran variedad de tests que se han desarrollado y han convertido a la Psicología en una ciencia positiva.

9. DESARROLLO DE LA TEORÍA DE LOS TESTS

Como consecuencia del auge conseguido por los tests surge la necesidad de desarrollar un marco teórico que sirva de fundamento a las puntuaciones obtenidas por los sujetos cuando se les aplican, posibilite la validación de las interpretaciones e inferencias realizadas a partir de ellas, y permita la estimación de los errores de medida inherentes a todo proceso de medición a través del desarrollo de una serie de modelos.

Partiendo de la base de que las características psicológicas que se intentan medir no son, por lo general, observables de forma directa, y considerando que los tests son los instrumentos de medida que se van a utilizar para llevar a cabo la medición de tales características, cuando se aplica un test a una muestra de sujetos se pueden plantear varias preguntas: ¿en qué medida esas puntuaciones reflejan el nivel de cada sujeto en la característica o rasgo medido?, ¿cómo estar seguros de que la puntuación obtenida es la que le corresponde a cada sujeto realmente? Si en todo proceso de medición, aunque sea en el campo de la física o de la biología, se cometen errores de medida, ¿cuántos más se cometerán cuando se trata de medir variables psicológicas?, ¿qué error está afectando a esas puntuaciones obtenidas al aplicar el test a la muestra de sujetos?

Ante estos problemas es lógico pensar en la necesidad de algún procedimiento que permita analizar, por una parte, la precisión de las medidas obtenidas; es decir, hasta qué punto las puntuaciones obtenidas por los sujetos en un test equivalen a sus puntuaciones verdaderas y cuál es la cuantía del error de medida que las afecta *(fiabilidad de las puntuaciones)* y, por otra parte, la validez de las inferencias o conclusiones que se puedan sacar a partir de las mismas (*validez*).

Para dar respuesta a estos problemas se desarrolló un marco teórico general, la *Teoría de los Tests,* que va a permitir establecer una relación funcional entre las variables observables (a partir de las puntuaciones empíricas obtenidas por los sujetos en los tests o en los ítems que los componen), y las variables inobservables (las puntuaciones verdaderas o el nivel de habilidad de los su-

jetos en el rasgo que se está midiendo) (Gulliksen, 1950; Lewis,1986 y Lord y Novick, 1968).

Para poder efectuar inferencias a partir de las puntuaciones de los sujetos en los tests, es necesario que la relación entre el nivel de rasgo, constructo o característica, que se quiere medir y las puntuaciones empíricas obtenidas pueda ser establecida a partir de una función matemática o modelo. Cada uno de estos modelos representa un tipo de relación funcional, y mediante una serie de supuestos deberá especificar los factores que influyen en las puntuaciones obtenidas por los sujetos en los tests. En la medida en que los supuestos sean válidos, las inferencias lógicas (matemáticas) que se realicen a partir del modelo describirán de forma correcta las propiedades de las puntuaciones de los tests, en caso contrario estas inferencias serán incorrectas.

Cada modelo podría dar origen a una Teoría de los Tests, pero las que han tenido una mayor incidencia en este campo han sido: la *Teoría Clásica de los Tests* (TCT), y la *Teoría de Respuesta al Ítem* (TRI).

Nota: Dado que en los estudios del Título de Grado de Psicología en la UNED, la Psicometría es una asignatura cuatrimestral, los contenidos de estas este texto se centrarán en la Teoría Clásica de los Tests. Aquellos que lo deseen podrán completar su formación psicométrica en los estudios de Postgrado.

9.1. La Teoría Clásica de los Tests (TCT)

La TCT se desarrolló, fundamentalmente, a partir de las aportaciones de Galton, Pearson y Spearman, y gira en torno a tres conceptos básicos: las puntuaciones empíricas u observadas (X), las puntuaciones verdaderas (V) y las puntuaciones debidas al error (E). Las primeras corresponden a las puntuaciones obtenidas por los sujetos cuando se les aplica un test, las puntuaciones verdaderas son las que realmente tienen los sujetos en el rasgo o constructo medido y coincidirían con las empíricas en el caso hipotético de que no existieran los errores de medida (E).

Para establecer la relación funcional entre esos tres conceptos la TCT se sirve del modelo más simple, *el modelo lineal*. Este modelo fue desarrollado por Spearman y formulado en una serie de trabajos fechados en 1904, 1907, 1910 y 1913. Para una revisión sistemática de la TCT es necesario acudir a la obra de Guilford (1954) *Psychometric Methods* y sobre todo a la de Gulliksen (1950) *Theory of Mental Tests*.

El modelo lineal de Spearman, cuyos supuestos serán desarrollados en el Tema 4 de este libro, es un modelo aditivo en el que la puntuación observada (variable dependiente) de un sujeto en un test (X) es el resultado de la suma de dos componentes: su puntuación verdadera (variable independiente) (V) y el error (E) que inevitablemente lleva asociado todo proceso de medición. A partir de los supuestos del modelo y de las deducciones que se extraigan a partir de los mismos, se podrá hacer una estimación de esos errores. La expresión formal del modelo es:

$$X = V + E \qquad [1.5]$$

La ejecución de un sujeto al responder a un test en un momento determinado estará afectada por múltiples factores difícilmente controlables, lo que implicará que la puntuación obtenida, su puntuación empírica, no coincida con su puntuación verdadera. Ante la imposibilidad de saber con exactitud cual es esta puntuación verdadera será necesario hacer estimaciones de la misma en base a los supuestos del modelo.

A pesar de que en el modelo lineal de Spearman sólo se encuentra un término de error en el que se incluirían todos los errores aleatorios que están afectando a las puntuaciones empíricas u observadas, los errores pueden provenir de numerosas fuentes: una de las fuentes de error puede ser el mismo sujeto ya que cualquier cosa que le haya ocurrido, su estado emocional, su cansancio y fatiga, etc., puede estar incidiendo en el rendimiento en el test y, por lo tanto, en la puntuación que obtenga; otra fuente de error puede provenir del propio test debido a los ítems que lo forman y al tipo de formato; también las características de los aplicadores del test pueden estar incidiendo en la puntuación de los sujetos en el mismo; otras fuentes pueden ser las condiciones ambientales y las instrucciones que se den, etc.

Aunque resulta imposible separar inequívocamente cuales son los factores que contribuyen a la puntuación verdadera y los que contribuyen al error (Feldt y Brennan, 1989), se han realizado algunos intentos para sistematizar y clasificar el error en función de las posibles fuentes que lo originan (Bock y Wood, 1971; Novick, 1966; Sutcliffe, 1965; Stanley, 1971; Thorndike, 1951,1989) y se han propuesto algunos modelos que suponen variaciones o extensiones del modelo de Spearman. Las variaciones encontradas entre estos modelos alternativos se deben a distintas matizaciones hechas respecto a los errores. Sin embargo el más ambicioso y global de los intentos realizados para estimar la fiabilidad de un instrumento de medida, analizando de forma sistemática las posibles fuentes de error es el proporcionado por la *Teoría de la Generalizabilidad (TG)* propuesta por Cronbach y sus colaboradores (Glesser, Cronbach y Rajaratnam, 1965; Cronbach, Rajaratnam, Glesser, 1963; Cronbach, Glesser, Nanda y Rajaratnam, 1972) que tiene en cuenta todas las posibles fuentes de error (las debidas a factores individuales, situacionales, características del evaluador y variables instrumentales) e intenta diferenciarlas mediante la aplicación de los procedimientos clásicos de análisis de varianza (AVAR). Las fuentes de variación (factores en términos de ANOVA) se denominan facetas y los niveles de cada factor condiciones. La medida psicológica se convierte en un índice obtenido en una muestra y el problema, como en toda inferencia, será generalizar esa medida. En castellano pueden consultarse Martínez-Arias (1995) o Paz-Caballero (1992).

A pesar del avance que en cierto modo supuso la TG, suele ser considerada más como una extensión de la TCT que como un modelo alternativo. Por otra parte, la complicación de sus diseños y la aparición de nuevos modelos psicométricos englobados bajo la denominación de *Teoría de*

Respuesta al Ítem (TRI) fueron la causa de que su utilización en el marco de la Teoría de los Tests quedara relegada a segundo término.

9.2. Teoría de Respuesta al Ítem (TRI)

Lord (1953) observó que cuando a una muestra de sujetos se les aplicaba un test, o una serie de tests, para evaluar su nivel en un determinado rasgo, la puntuación obtenida dependía del conjunto de ítems o tests utilizados cuando, en realidad, su nivel en el rasgo en el momento de la aplicación no tenía porqué variar. Los sujetos no debían tener puntuaciones altas o bajas en un test en función de que los ítems que lo formaran fueran más fáciles o difíciles. También los estadísticos de los ítems, su índice de dificultad y de discriminación, dependían de la muestra de sujetos utilizada para su cálculo.

Estos dos problemas, junto con el de la indiferenciación del error (sólo había un componente error que englobaba a todos), fueron los que centraron las críticas hechas a la TCT. Los intentos para solucionar el último de ellos ya han sido comentados; para intentar solucionar los dos restantes algunos psicómetras, entre los que se puede citar a Gulliksen (1950) y el mismo Lord (1952, 1953), se interesaron en el desarrollo de teorías y modelos que permitieran describir los niveles de habilidad de los sujetos con independencia de la muestra de ítems o de tareas utilizados para su evaluación, y el cálculo de los estadísticos de los ítems con independencia de la muestra de sujetos utilizada. La solución más adecuada se encontró en el marco de la *Teoría de Respuesta al Ítem (TRI)*, que proporciona una serie de modelos que asumen una relación funcional entre los valores de la variable que miden los ítems (nivel de habilidad de los sujetos en el rasgo medido) y la probabilidad de que los sujetos, en función de su nivel de habilidad, acierten cada ítem. A esta función se la conoce con el nombre de *Curva Característica del Ítem* debido a que, realmente, es la curva que caracteriza a cada uno de ellos. La probabilidad de que un sujeto acierte a cada uno de los ítems no depende ya del propio ítem depende, exclusivamente, del nivel de los sujetos en la variable que mide cada uno de ellos.

En 1952, Lord defendió su tesis doctoral en la que presentó a la TRI como un modelo o teoría con entidad propia, de ahí que sea considerado el padre y fundador de la TRI. Como resultado de su tesis se publicó en *Psychometric Monographs* nº 7, una monografía bajo el título *A Theory of Test Scores*, a este trabajo siguieron otros que marcaron el comienzo de una nueva manera de trabajar en el campo de la Psicometría (Birnbaum, 1957, 1958a, 1958b; Lord y Novick, 1968; Rasch, 1960).

El desarrollo de estos modelos supuso un gran avance en la Teoría de los Tests; sin embargo, la dificultad de utilizarlos en la práctica sin la ayuda de los ordenadores fue la causa de que su gran desarrollo no llegara hasta finales del siglo XX, cuando ya el uso de los ordenadores personales fue habitual y asequible para una gran mayoría y, además, se desarrollaron los programas de software necesarios para su utilización.

A pesar del gran desarrollo de la TRI hacia 1980, la TCT sigue en auge ya que hay problemas que se pueden solucionar más eficaz y rápidamente dentro de este marco.

Nota: El lector interesado en la Teoría de Respuesta al Ítem puede consultar, en castellano, los siguientes textos introductorios: Martínez-Árias, M.R. (1995) *Psicometría: Teoría de los tests psicológicos y educativos*. Madrid: Síntesis (capítulos 10 y 11); Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems*. Madrid: Pirámide; Santisteban, C. (1990/1995). *Psicometría. Teoría y práctica en la construcción de tests*. Madrid: Norma (capítulos 10 y 14).

10. EJERCICIOS DE AUTOEVALUACIÓN

A continuación se proponen una serie de afirmaciones que pueden ser correctas o incorrectas; el lector, después de leerlas detenidamente, deberá responder en un sentido o en otro razonando su respuesta.

1. La Teoría de la Medición es uno de los campos de la Psicometría.
2. Las dos vías a través de las cuales se desarrolló la Psicometría fueron los estudios de Psicofísica y las investigaciones acerca de las diferencias individuales.
3. Los métodos psicofísicos se desarrollaron para su utilización en el campo de la percepción.
4. A través de los métodos de escalamiento psicológico se pueden medir variables que no tengan ninguna dimensión física subyacente.
5. Los métodos psicofísicos están vinculados a la Psicología correlacional.
6. Los tests de Galton medían funciones mentales.
7. Los atributos psicológicos son variables directamente observables.
8. Los métodos de escalamiento psicofísico pretenden ordenar a los sujetos a lo largo de un continuo psicológico.
9. Cuando hay varias dimensiones subyacentes a los estímulos, éstos reciben un valor escalar en cada una de ellas.
10. La Psicometría utiliza como método de investigación el método científico.
11. Podemos considerar a los tests como instrumentos de medición.
12. Spearman desarrolló un modelo para las puntuaciones de los tests.
13. Un problema fundamental en la medición psicológica es el del control del error de medida.
14. El cociente intelectual es una norma cronológica que permite la ordenación de los sujetos.
15. Binet fue el primero en considerar la importancia de los procesos mentales superiores en el estudio de las diferencias individuales.
16. Las puntuaciones obtenidas por los sujetos en un test referido al criterio se comparan con las obtenidas por un grupo normativo.
17. El objetivo de los tests referidos a las normas es poner de manifiesto las diferencias individuales en el rasgo que miden.
18. Los Tests Adaptativos Informatizados (TAI s) son los mismos tests de papel y lápiz pero aplicados por ordenador.

19. Una de las críticas a la Teoría Clásica de los Tests es que los parámetros de los ítems dependen de la muestra de sujetos a los que se les aplican y las puntuaciones de los sujetos en el rasgo medido dependen de los ítems a los que responden.
20. Los *métodos directos* de elaboración de escalas psicofísicas utilizan una escala de respuesta.
21. Los *métodos directos* se utilizan para la obtención de umbrales.
22. El umbral absoluto marca el origen de la escala de sensación.
23. El paso de la sensación a la no sensación viene determinado por el umbral diferencial.
24. En el modelo escalar de Thurstone los estímulos se ordenan a lo largo de un continuo físico.

11. SOLUCIONES A LOS EJERCICIOS DE AUTOEVALUACIÓN

1. La afirmación es correcta.

   Hemos considerado a la Psicometría como una disciplina, dentro del área de la Psicología, que tiene unas funciones concretas, tanto en lo que se refiere a sus implicaciones teóricas como prácticas. En este sentido, la Teoría de la Medición tiene como objetivo legitimar y justificar la medición de variables psicológicas dándole una fundamentación teórica.

2. La afirmación es correcta.

   Como hemos comentado anteriormente, las dos vías de desarrollo de la Psicometría fueron los estudios de Psicofísica y las investigaciones acerca de las diferencias individuales. Aunque en un principio pudiera parecer que ambas vías de desarrollo siguieron caminos distintos, podemos comprobar la incidencia que tienen los avances logrados en cada una de ellas sobre la otra.

3. La afirmación es correcta.

   Los estudios de Psicofísica tratan de relacionar la magnitud de los estímulos con la percepción que de ellos tienen los sujetos; es decir, con la sensación que les producen.

4. La afirmación es correcta

   Así como los métodos psicofísicos se utilizan para estudiar las relaciones entre un conjunto de estímulos que varían a lo largo de un continuo físico y las sensaciones que producen en los sujetos que varían a lo largo de un continuo psicológico, como ocurre, por ejemplo, si queremos establecer una correspondencia entre la intensidad de luz de una serie de estímulos y la sensación de luminosidad que producen; si tratamos de medir las sensaciones que producen en una muestra de sujetos una serie de obras de arte, la carga de violencia de determinadas escenas de películas, etc., nos encontramos con que los estímulos no varían a lo largo de ningún continuo físico sino psicológico y las sensaciones que producen en los sujetos también varían a lo largo de un continuo psicológico. Por eso, para la medición de este tipo de variables psicológicas se utilizan otros métodos de escalamiento, los métodos psicológicos, cuyo principal impulsor fue Thurstone.

5. La afirmación es incorrecta.

   Los métodos psicofísicos están vinculados a la Psicología experimental; son los estudios de las diferencias individuales los que están vinculados a la Psicología correlacional.

6. La afirmación es incorrecta.

   Los tests de Galton, que pueden ser considerados los primeros tests psicométricos, tenían un marcado carácter sensorial y motor. Aunque Galton pensaba que los datos antropométricos que estaba reuniendo para sus investigaciones le podían valer para estudiar las dimensiones de la mente del hombre, al final de sus días debió admitir que esas medidas no tenían valor como medidas de la inteligencia.

7. La afirmación es incorrecta.

   Las variables psicológicas o atributos psicológicos, son variables hipotéticas denominadas constructos que no se pueden medir de forma directa porque no son variables directamente observables. Es preciso determinar una muestra de conductas representativas del atributo que queremos estudiar y, dado que estas conductas son variables observables susceptibles de medición, a partir de ellas haremos inferencias acerca del atributo.

8. La afirmación es incorrecta.

   Por una parte los métodos psicofísicos no ordenan a los sujetos puesto que son métodos para escalar (ordenar) estímulos y, por otra, los métodos de escalamiento psicofísico permiten relacionar dos continuos, un continuo físico a lo largo del cual varían los estímulos y un continuo psicológico a lo largo del que se sitúan las sensaciones que producen los estímulos.

9. La afirmación es correcta.

   A diferencia de los métodos de escalamiento unidimensional en los que los estímulos reciben un valor escalar en la única dimensión que se mide, en los métodos de escalamiento multidimensional, al estar considerándose varias dimensiones a la vez, los estímulos recibirán un valor escalar en cada una de ellas.

10. La afirmación es correcta.

    La Psicología, como disciplina científica que es, utiliza el método científico para la adquisición del conocimiento, adaptándole a sus necesidades particulares.

11. La afirmación es correcta.

    El método de los tests se desarrolla para el estudio de las diferencias individuales y, para ello, desarrolla los tests como instrumentos de medición.

12. La afirmación es correcta.

    En efecto, Spearman desarrolló un modelo lineal que supuso un gran desa-rrollo para el método de los tests (Teoría Clásica de los Tests). Este modelo partía del supuesto de que la puntuación observada a través de los tests (X) está formada por dos componentes, una componente debida a la verdadera medida del rasgo (puntuación verdadera del sujeto en el rasgo) y otra una componente errónea (el error asociado a todo proceso de medición), y que la relación entre estas dos componentes es aditiva.

13. La afirmación es correcta.

    En todo proceso de medición, aún en los llevados a cabo en las ciencias exactas, se cometen errores de medición. Este problema se agrava cuando las características que se quieren medir

no pueden ser observadas directamente y han de ser medidas mediante indicadores. En este caso el control del error cometido es un problema fundamental.

14. La afirmación es correcta.

Se trata de un término acuñado por Stern (1912) y que equivale al cociente entre la edad mental y la edad cronológica, multiplicado por cien para evitar los decimales.

15. La afirmación es correcta.

Binet comprendió que las sensaciones no jugaban un papel demasiado importante en la Psicología diferencial y que había que recurrir al estudio de los procesos mentales superiores.

16. La afirmación es incorrecta.

Una de las diferencias entre el enfoque normativo y el de los tests referidos al criterio es precisamente que, en estos últimos, no se necesita un grupo normativo para la interpretación de las puntuaciones obtenidas por los sujetos, sino que éstas se interpretan en relación a un dominio de contenidos o conductas.

17. La afirmación es correcta.

Los resultados se interpretan en relación a los obtenidos por el grupo normativo.

18. La afirmación es incorrecta.

Una cosa son los tests aplicados por ordenador y otra los tests adaptativos informatizados. En éstos, los sujetos no tienen que contestar ni a los mismos ítems ni a todos los ítems de un tests. La selección de cada ítem se va haciendo de manera que se vayan *adaptando* al nivel de cada sujeto.

19. La afirmación es correcta.

En el marco de la TCT, los valores de los parámetros de los ítems dependen de la muestra de sujetos a los que se les han aplicado. Si el índice de dificultad de un ítem se obtiene calculando la proporción de sujetos que han acertado ese ítem, es fácil darse cuenta de que esa proporción variará en función del nivel de los sujetos. Por otra parte, el nivel de aptitud o habilidad de los sujetos depende de que los ítems a los que respondan sean más fáciles o difíciles. Este problema no tenía una solución real dentro del marco de la TCT y hubo que esperar al desarrollo de la TRI para que se pudiera solucionar.

20. La afirmación es correcta.

En eso se diferencian de los métodos indirectos asociados a la psicofísica de Fechner, ya que éstos utilizan una escala de sensación elaborada a base de ir sumando las diferencias apenas perceptibles (dap).

21. La afirmación es incorrecta.

Los métodos directos no implican el cálculo de umbrales, en estos métodos el sujeto emite de forma directa su respuesta.

22. La afirmación es correcta.

El umbral absoluto es el valor mínimo que tiene que tener un estímulo para poder ser percibido por el sujeto. Este valor mínimo en la escala física se empareja con el valor cero de la escala de sensación y, por lo tanto, marca su origen.

23. La afirmación es incorrecta.

El paso de la sensación a la no sensación (o viceversa) equivale, en la escala física, al umbral absoluto. El umbral diferencial es el incremento mínimo que tiene que experimentar la magnitud de un estímulo para que el sujeto perciba que ha habido un cambio.

24. La afirmación es incorrecta.

Precisamente la gran aportación de Thurstone fue elaborar un modelo de escalamiento en el que no fuera necesario recurrir a ningún continuo físico.

12. BIBLIOGRAFÍA COMPLEMENTARIA

Barbero, M.I. (2007). *Métodos de elaboración de escalas.* Madrid: UNED.

A lo largo del libro los alumnos podrán encontrar información sobre algunos de los principales métodos de escalamiento

Muñiz, J. (1998, 2008). *Teoría Clásica de los tests.* Madrid: Pirámide

En el primer capítulo se hace una buena introducción sobre el origen y desarrollo de los tests y de la teoría de los tests.

Parte I

CONSTRUCCIÓN DE INSTRUMENTOS DE MEDICIÓN PSICOLÓGICA

TEMA 2

PRINCIPIOS BÁSICOS PARA LA CONSTRUCCIÓN DE INSTRUMENTOS DE MEDICIÓN PSICOLÓGICA

María Isabel Barbero García

SUMARIO

1. ORIENTACIONES DIDÁCTICAS

A lo largo del tema anterior se ha intentado dar a conocer a los alumnos lo que es la Psicometría y el papel que juega en el marco de la Metodología de las Ciencias del Comportamiento. Se ha puesto de manifiesto la dificultad que conlleva el intentar medir variables psicológicas y la forma de llevar a cabo el proceso de medición.

Después de exponer, aunque de forma muy esquemática, las dos vías de desarrollo de la Psicometría: la de los estudios de Psicofísica y la de los estudios de las Diferencias Individuales, se hizo una exposición de los distintos tipos de escalamiento según que el objeto a escalar fueran los estímulos, los sujetos o ambos y, posteriormente, se hizo una breve introducción a los principales métodos de escalamiento de estímulos y a los orígenes y desarrollo de los tests como instrumentos que van a permitir la medición de las diferencias individuales y, por lo tanto, el escalamiento de los sujetos. Asimismo, se expusieron las principales teorías que han ido surgiendo a lo largo de los años para justificar y legitimar las medidas obtenidas al aplicar los tests dándolas una fundamentación teórica.

Después de ese primer tema introductorio, en éste y a lo largo de los temas siguientes, y desde el marco de la Teoría Clásica de los Tests, vamos a centrarnos en todos aquellos aspectos relacionados con la construcción, evaluación y aplicación de los instrumentos de medición que van a permitir, entre otras cosas, el estudio de las diferencias individuales respecto a alguna característica psicológica, tomar decisiones acerca de si los alumnos han conseguido unos objetivos curriculares mínimos, detectar problemas comportamentales, etc. (ver tabla 2.2, página 62).

Trataremos de proporcionar a nuestros alumnos una herramienta de trabajo que les permita, en un momento determinado, no sólo poder construir un instrumento de medición psicológica y evaluar su calidad métrica, sino poder interpretar los resultados obtenidos con su aplicación y hacer inferencias y tomar decisiones en función de ellos.

Sé que para muchos se trata de una asignatura difícil cuya utilidad no descubren hasta el momento de la práctica profesional; pero, cuando llega ese momento, echan mano de *los viejos apuntes* para poder moverse con cierta soltura en algunas situaciones.

Se ha intentado utilizar un lenguaje que pudiera ser comprendido por todas aquellas personas que quisieran utilizar este libro como guía en su formación dentro de nuestro campo. Aunque la mayoría de nuestros alumnos no se van a dedicar a la construcción de tests, y por lo tanto tienen un interés relativo por el tema que nos ocupa, es cierto, y lo he constatado a lo largo de los muchos años de experiencia, que todos ellos se van a enfrentar en algún momento de su vida profesional con los tests y es necesario que, con una mayor o menor profundidad, tengan un cierto conocimiento de lo que tienen entre manos.

Para poder utilizar un test como instrumento científico de medición es necesario cubrir una serie de etapas. Una etapa inicial, que abarcaría la elaboración de una prueba piloto, su aplicación a una muestra de sujetos y la asignación de puntuaciones a los mismos; una segunda etapa en la que se evaluaría la calidad psicométrica de cada uno de los ítems que componen la prueba piloto así como del conjunto de la misma, y finalizaría con la construcción definitiva del test, y una tercera etapa en la que se llevaría a cabo la aplicación del test a una muestra representativa de la población a la que va dirigido, se asignarían puntuaciones a los sujetos para su evaluación y se llevaría a cabo el proceso de estandarización de las puntuaciones obtenidas y el establecimiento de normas que permitan su interpretación. El proceso finalizaría con la elaboración del manual del test que deberá incluir toda la información necesaria para que pueda ser utilizado por los psicólogos especializados.

En este tema se va abordar la primera etapa, la etapa inicial que incluye la elaboración de una prueba piloto y su aplicación a una muestra de sujetos, y en los temas siguientes se abordará el estudio del resto de las etapas.

Los puntos básicos en los que han de profundizar los alumnos a la hora de abordar el estudio de este tema son:

— Tener muy claras las distintas fases que hay que cubrir a la hora de construir la prueba piloto de un test: definición de la finalidad del test, especificación de las características del test, redacción de los ítems, revisión por un grupo de expertos y confección de la prueba piloto.

— Conocer las distintas decisiones que hay que tomar en cada una de las fases y la forma de hacerlo:

a) En la fase de definición de la finalidad del test hay que tomar decisiones acerca de qué constructo o variable psicológica se va a medir, a qué población va a ir destinado el test y cuál va a ser la utilización prevista.

b) En la fase de especificación de las características del tests se debe decidir sobre el contenido del test, qué tipo de formato van a tener los ítems, cuál va a ser la longitud de la prueba y qué características psicométricas son las más adecuadas.

c) Es necesario que los alumnos conozcan qué tipo de formato es el más adecuado en función de la variable psicológica que se quiera medir y las reglas que se deben seguir, sean

éstas de carácter general o específicas de cada tipo de formato, para que la redacción de los ítems sea lo más correcta posible.

d) Una vez redactados los ítems se ha de tomar la decisión acerca de qué persona/s han de hacer una revisión crítica de los mismos para conseguir una mejor calidad.

e) La confección de la prueba piloto implica tomar decisiones acerca de las instrucciones que se van a incluir, cuál va a ser la forma de administrar la prueba y de qué manera se van a registrar las respuestas de los sujetos

Aunque a lo largo del texto se han incluido varios ejemplos, creemos necesario que el alumno realice también los ejercicios de autoevaluación; de esta manera podrá él mismo controlar su aprendizaje.

2. LOS TESTS, ESCALAS, CUESTIONARIOS E INVENTARIOS

En la literatura científica los instrumentos utilizados para medir variables psicológicas han recibido distintas denominaciones: Tests, Escalas, Cuestionarios, Inventarios, etc., y no siempre ha quedado claro el porqué de esta distinción; es cierto que no siempre es posible diferenciar estos términos puesto que se han utilizado indistintamente, sobre todo algunos; pero vamos a tratar de clarificarlos o, por lo menos, clarificar el sentido que van a tener dentro de este libro.

En general, se ha utilizado el término *Test* como un término general para referirnos a cualquier instrumento de medición psicológica, es el término más utilizado, pero sobre todo se utiliza para hacer referencia a los instrumentos de medición de variables de tipo cognitivo: aptitudes, conocimientos, rendimiento, habilidades, etc., en los que las respuestas de los sujetos a cada uno de los elementos que los forman son correctas o incorrectas y la puntuación total en el test viene dada por la suma de todas las respuestas correctas obtenidas (bien suma directa o ponderada), dando lugar a una escala jerárquica o acumulativa.

El término *Escala* se suele utilizar para hacer referencia a los instrumentos elaborados para medir variables no cognitivas: actitudes, intereses, preferencias, opiniones, etc., y se caracterizan porque los sujetos han de responder eligiendo, sobre una *escala de categorías graduada y ordenada,* aquella categoría que mejor represente su posición respecto a aquello que se está midiendo, no hay respuestas correctas o incorrectas, y la puntuación total de los sujetos en la escala será la suma de las puntuaciones asignadas a las categorías elegidas por los sujetos.

Los *Cuestionarios* suelen están formados a menudo por una serie de ítems o elementos que no están necesariamente relacionados unos con otros, cuyas *opciones de respuesta no están ordenadas ni graduadas,* que pueden ser puntuados e interpretados individualmente y en los que tampoco

hay respuestas correctas o incorrectas. Las preguntas que incluyen pueden ser muy variadas con el fin de obtener una mayor información acerca del sujeto/s y de su entorno (edad, profesión, nivel de ingresos, nivel de estudios, opiniones acerca del problema que se quiere evaluar, etc). Son el instrumento generalmente utilizado cuando se quiere llevar a cabo una investigación mediante encuestas.

El término *Inventario* suele estar vinculado a los instrumentos elaborados para medir variables de personalidad. Las respuestas de los sujetos a los ítems no son correctas o incorrectas, lo único que demuestran es la conformidad o no de los sujetos con los enunciados de los ítems.

EJEMPLOS:

1. La capital de España es:
   a) París
   b) Londres
   c) Madrid
2. Debería estar prohibido fumar en todos los restaurantes:
   1) Completamente de acuerdo
   2) De acuerdo
   3) Indiferente
   4) En desacuerdo
   5) Completamente en desacuerdo
3. ¿Qué parte de la asignatura de Psicometría le resulta más fácil de estudiar?
   a) Fiabilidad
   b) Validez
   c) Análisis de elementos
   d) La construcción de tests
4. A menudo me siento como si los demás me ignoraran V F
   El problema de mucha gente es que no toma las cosas en serio V F
   Creo que me gustaría trabajar en unos grandes almacenes V F

El primer ejemplo representa una pregunta típica de lo que denominamos un *Test de Conocimientos*, el sujeto ha de elegir la respuesta correcta entre las distintas alternativas que se le ofrecen. El segundo ejemplo podría ser una pregunta típica de una *Escala* que midiera la actitud ante el tabaquismo. Para cada elemento se ofrece una escala de respuesta graduada, en este caso del 1 al 5,

de manera que la opción elegida por los sujetos refleje su posición personal ante el enunciado del ítem. No hay respuestas correctas o incorrectas. El tercer ejemplo, sería una pregunta típica de un *Cuestionario*, tampoco hay respuestas correctas o incorrectas, simplemente el sujeto deberá elegir la alternativa que refleje su opinión; pero, a diferencia de las alternativas presentadas en las *Escalas*, en las que el sujeto debía elegir entre una serie de categorías ordenadas en función del grado de acuerdo con respecto al enunciado del ítem, en los *Cuestionarios* las distintas opciones no guardan ninguna relación de orden entre sí, de manera que es indiferente su orden de presentación. Finalmente, los enunciados incluidos en el ejemplo cuarto podrían ser típicos de un inventario de personalidad, como se puede observar no hay respuestas correctas o incorrectas, los sujetos deben leer cada uno de los enunciados y responder si es verdadero (o habitualmente verdadero) o falso (o habitualmente falso) para ellos; es decir, su acuerdo o no con los distintos enunciados, aunque en los inventarios de personalidad al no haber respuestas correctas o incorrectas se suelen utilizar otras etiquetas lingüísticas.

EN RESUMEN:

El término *Test* suele reservarse, generalmente, para todos aquellos instrumentos de medición utilizados en el ámbito cognitivo; es decir, para aquellos instrumentos construidos para medir: aptitudes, conocimientos, rendimiento, habilidades, etc.

Cuando se quieren medir variables que no pertenecen al ámbito cognitivo, sino al afectivo u oréctico (personalidad, actitudes, intereses, opiniones, etc.), se utilizan las *Escalas,* los *Cuestionarios* y los *Inventarios*. Las escalas muestran el enunciado del elemento junto a una serie de alternativas de respuesta categorizadas de forma ordenada (escalas de categorías o escalas de clasificación), de manera que el sujeto pueda mostrar su posición respecto a lo que se muestra en el enunciado, eligiendo la categoría con la que se sienta más identificado. Los cuestionarios suelen estar formados por elementos cuyas alternativas de respuesta no forman categorías ordenadas (listados). Cada alternativa es independiente de las demás. En este caso, la tarea del sujeto consiste en elegir la alternativa o alternativas que prefiera o la/s que, en su opinión, refleje mejor aquello que se está valorando, y los inventarios se suelen reducir a una serie de enunciados ante los que los sujetos han de responder en función de su grado de acuerdo o desacuerdo.

A lo largo del texto se va a utilizar la palabra *Test* para hablar en términos generales de todos los instrumentos de medición psicológica ya que es el término más amplio y aceptado internacionalmente, y se utilizarán los otros términos cuando se haga referencia expresa a un determinado tipo de instrumentos de medición.

3. EL PROCESO DE CONSTRUCCIÓN DE UN TEST

La construcción de un test es un proceso laborioso que requiere, como paso previo al proceso de construcción propiamente dicho, tener muy claro qué se quiere medir con él, a quién va a ir dirigido y para qué se va a utilizar. El conocimiento del objetivo del test es el paso previo que va a condicionar las siguientes etapas de la construcción (Crocker y Algina, 1986; Millman y Greene, 1989).

La tabla 2.1 recoge de forma esquemática las distintas etapas que hay que ir cubriendo desde el momento en que el investigador se plantea la tarea de construir un test hasta el momento de la aplicación piloto de la prueba desarrollada, recogiendo todos aquellos aspectos a los que hay que dar respuesta. No se incluyen las etapas correspondientes a la evaluación de las características psicométricas de la prueba, ni a las de la aplicación del test definitivo, porque se analizarán en los temas siguientes tal y como se comentó al principio.

TABLA 2.1.
Esquema de las etapas a cubrir para la construcción de una prueba piloto

Etapas	Decisiones a tomar acerca de:
■ Determinar la finalidad del test	■ Qué se va a medir ■ A quien se va a medir ■ Para qué se va a medir
■ Especificar las características del test	■ Cuál va a ser el contenido ■ Qué tipo de ítems se van a incluir ■ Cuántos ítems ■ Características psicométricas
■ Redacción de los ítems	■ Ítems de elección ■ Ítems de construcción
■ Revisión crítica de los ítems por un grupo de expertos	■ Qué ítems van a seleccionarse
■ Confección de la prueba piloto	■ Instrucciones de administración ■ Formato de presentación ■ Formato de registro de respuestas
■ Aplicación de la prueba piloto	■ Individual – colectiva ■ Papel y lápiz- Informatizada ■ Correo, mediante entrevista personal, por teléfono, etc.
■ Corrección de la prueba piloto y asignación de puntuaciones a los sujetos	■ En los tests formados por ítems de elección ■ En los tests formados por ítems de construcción

4. LA FINALIDAD DEL TEST

Como paso previo a la construcción de un test es necesario tener muy claro cual va a ser su objetivo; es decir qué es lo que se quiere medir, a quién se quiere medir y para qué se quiere medir.

Supongamos que los profesores de Psicometría queremos construir un test (examen) de Psicometría para evaluar los puntos fuertes y débiles del conjunto de los alumnos respecto al programa de nuestra asignatura y poder incluir en la página Web algunas recomendaciones que les permitan superar las dificultades encontradas durante el estudio. Ya tenemos el objetivo del test:

■ ¿Qué se quiere medir?	Los conocimientos de Psicometría
■ ¿A quién se quiere medir?	A los alumnos
■ ¿Para qué?	Para detectar puntos fuertes y débiles en los alumnos respecto al programa de la asignatura.

La primera pregunta hace referencia a la variable objeto de estudio, aquella que se quiere medir; la segunda a la población a la que va dirigido y la tercera al uso o utilidad que se va a dar al test.

Otro ejemplo podría ser el siguiente:

Un grupo de Psicólogos, especialistas en modificación de conducta, encargan a un grupo de psicómetras la construcción de un test que permita medir el nivel de estrés que producen los exámenes de acceso a la Universidad con el fin de buscar algún sistema que permita reducirlo o, al menos, controlarlo. La variable a medir sería el estrés, la población objeto de estudio estaría formada por todos los alumnos que realizan el examen de acceso a la Universidad, el objetivo sería buscar algún sistema para reducir el nivel de estrés o, al menos, controlarlo.

4.1. La variable objeto de estudio

Quizás pueda parecer una trivialidad el hacer tanto hincapié en la necesidad de conocer claramente lo que se quiere medir antes de iniciar la construcción de un test; a lo mejor lo sería si la variable a medir fuera una variable de tipo físico, como la longitud, la altura, etc., ya que son variables que están bastante claras para todo el mundo y todos saben cómo hay que medirlas. El problema surge cuando lo que se trata de medir es una variable psicológica, inobservable de forma directa; una variable o atributo latente a la que se le da también la denominación de *constructo*.

Los constructos (variables teóricas inobservables), tal y como ya hemos expuesto anteriormente, se manifiestan a través de una serie de conductas que ya sí son observables de forma directa y, por lo tanto, sí son susceptibles de medición. Estas conductas, para que puedan ser consideradas como manifestaciones del constructo han de ser más o menos uniformes y constantes a lo largo del tiempo y en una serie de situaciones. Así, por ejemplo, diremos que una persona es *agresiva*, no porque en una ocasión haya dado muestras de agresividad, sino porque cuando se enfrenta a ciertas situaciones lo normal es que reaccione siempre agresivamente. Ahora bien, ¿en qué consiste dar muestras de agresividad o reaccionar de forma agresiva?, hay una serie de conductas que pueden ser consideradas como tales; por ejemplo, un niño tendrá un comportamiento agresivo si en los recreos pelea con los demás niños sin motivo, si cuando se le regaña reacciona violentamente rompiendo alguna cosa, etc.

Todas las cuestiones que hacen referencia a estas conductas son las que deben ser reflejadas en los ítems del test; de ahí la importancia de definir claramente el constructo que se quiere medir puesto que, en la medida en que el constructo esté mejor definido, se tendrá un mayor y mejor conocimiento de las conductas en las que se manifiesta, evitándose la omisión de algunas áreas de conducta importantes y la inclusión de otras que resulten irrelevantes.

Por ejemplo, si quisiéramos construir un test para medir la *habilidad verbal* o la *impulsividad* lo primero que tenemos que hacer es definir ambos constructos. Una definición puede ser la propuesta por Thorndike (1989):

- *Habilidad verbal:* Se trata de un constructo que se pone de manifiesto por el conocimiento de significados de un gran número de palabras y elección de la palabra más adecuada en un determinado contexto.
- *Impulsividad:* Constructo que se manifiesta en una tendencia a autodescribirse con frases que implican decisiones rápidas y precisas para actividades no planificadas, nacidas del momento, a diferencia de las frases autodescriptivas que implican deliberación, tendencia hacia la planificación cuidadosa y reprimida.

Si el constructo está bien definido será más fácil determinar cuales son las conductas representativas del mismo y, a partir de ellas, especificar el contenido del test.

4.2. Población a la que va dirigido

Este es otro punto a tener en cuenta a la hora de construir un test, no es lo mismo construir un test para evaluar algún rasgo o característica en una población infantil que en una población de adultos, el contenido de los ítems, su redacción, la longitud del test y las instrucciones de aplicación y cumplimentación, por ejemplo, serán diferentes según el tipo de población a la que vaya dirigido.

Si se quiere construir un test para evaluar el razonamiento numérico, el contenido no podrá ser el mismo si se va a utilizar en alumnos del primer ciclo de enseñanza básica que si se va a utilizar en alumnos de bachillerato. Los contenidos del test deberán adecuarse al nivel de los alumnos.

Si se quiere evaluar algún rasgo de personalidad, el contenido de los ítems deberá adaptarse también a la población a la que va dirigido. Aunque el mismo constructo pueda manifestarse a través de algunas conductas comunes sea cual sea la población de interés, hay otro tipo de conductas, específicas de cada población, que son las que hay que resaltar.

En el caso de que se quisiera construir un test que midiera depresión, a la hora de buscar las conductas a través de las cuales se manifiesta el constructo se observaría que algunas de ellas son comunes a los niños y a los adultos, pero hay otras conductas, específicas de cada población, que son fundamentales para la evaluación del rasgo y que deberán quedar reflejadas en el test.

4.3. Utilización prevista

Además de tener en cuenta la población a la que va dirigido el test, otro de los aspectos fundamentales a tener en cuenta es la utilización que se le vaya a dar; es decir, para qué se va a utilizar, qué decisiones se van a tomar a partir de las puntuaciones que obtengan los sujetos. Un test puede utilizarse para tomar decisiones diferentes. Por ejemplo, una prueba de inglés puede utilizarse para evaluar el nivel medio de los alumnos en primero de bachillerato, para detectar aquellos alumnos que muestran un nivel deficiente, o puede utilizarse en una academia de idiomas para clasificar a los alumnos según su nivel de inglés y asignarlos a distintos cursos.

Por ejemplo, no es lo mismo querer construir un test de inteligencia general para detectar niños superdotados, que querer construirlo para detectar niños deficientes. En el primer caso, los ítems del test deberán ser en su mayoría muy difíciles, para que sólo puedan ser respondidos correctamente por los niños más inteligentes; mientras que, en el segundo caso, deberán ser muy fáciles, de manera que sólo sean respondidos de forma incorrecta por aquellos niños que muestren una inteligencia deficiente.

La tabla 2.2 muestra los usos más frecuentes de los tests así como las decisiones que se suelen tomar a partir de las puntuaciones obtenidas por los sujetos.

Del conocimiento de la finalidad del test; es decir, de las respuestas a qué se va a medir, a quién se va a medir y para qué se va a medir, van a depender las especificaciones que se deban hacer en cuanto al contenido del test, la dificultad de sus ítems, la longitud de la prueba, el tiempo de aplicación, etc. Especificaciones que iremos ampliando a continuación.

TABLA 2.2
Principales usos de los tests (adaptada de Nunnally, 2001, pág. 10)

Funciones	Decisiones a tomar
SELECCIÓN	*Ámbito educativo*: Se pueden utilizar para tomar decisiones acerca de si los alumnos han conseguido los objetivos curriculares mínimos y se les considera Aptos/No aptos. Por ejemplo el examen de acceso a la Universidad. *Ámbito profesional*: Se pueden utilizar para llevar a cabo la selección de un grupo de aspirantes a un determinado puesto de trabajo.
CLASIFICACIÓN/COLOCACIÓN	Se pueden utilizar en cualquier ámbito. Por ejemplo en el *ámbito educativo* se pueden utilizar para averiguar el nivel de los alumnos al principio de curso y asignarles a diferentes grupos en función de su nivel para que tengan un mejor aprovechamiento. Por ejemplo su nivel de Inglés.
DIAGNÓSTICO	Sea cual sea su ámbito de aplicación, los tests se pueden utilizar para detectar problemas comportamentales, de aprendizaje, etc. Para, una vez detectados poder poner los medios para tratar de solucionarlos; por ejemplo, mediante algún programa de intervención.
CERTIFICACIÓN	Acreditar, en el ámbito profesional, la cualificación o competencia de las personas para el desarrollo de su profesión y, en el ámbito académico, por ejemplo, para garantizar que han alcanzado los conocimientos y habilidades necesarias para poder obtener la titulación académica correspondiente.
ORIENTACIÓN/CONSEJO	Detectar las capacidades e intereses de las personas para, examinando las distintas opciones que se les presentan a nivel educativo, profesional y personal, elegir aquella/s que mejor se adecuen a su perfil.
DESCRIPCIÓN/INFORMACIÓN	Describir el rendimiento académico tanto a nivel individual como colectivo. Obtener información acerca de la opinión pública sobre algún aspecto, político, social, etc. de interés.

5. ESPECIFICACIÓN DE LAS CARACTERÍSTICAS DEL TEST

En esta etapa de la construcción de un test juega un papel importante la habilidad del constructor para planificar y desarrollar las especificaciones del mismo. Si consideramos que un test no es más que una herramienta que nos permite obtener datos para la medición y evaluación de alguna característica o atributo psicológico (constructo), la medición no será válida, o la evaluación no tendrá ningún sentido, a menos que el test se adecue a su objetivo.

Hay cuatro aspectos fundamentales que hay que tener en cuenta a la hora de desarrollar las especificaciones del test: su contenido, el formato de los ítems que se van a incluir, la longitud del test y la forma de aplicación.

5.1. Contenido

Una vez que se ha definido claramente el constructo que se quiere medir, la especificación del contenido comienza determinando cual es su dominio de conductas; es decir, el conjunto de conductas a través de las cuales se manifiesta. Cuando ya está determinado el dominio de conductas ya se pueden tomar decisiones acerca del contenido del test.

Hay ocasiones en las que el constructo no está claramente definido y, por lo tanto, no se tiene una idea clara del conjunto de las áreas de conducta a través de las cuales se puede manifestar. En este caso se recomienda, como señala Martínez-Arias (1995):

a) *Hacer un análisis del contenido del constructo*: haciendo preguntas abiertas a los sujetos acerca del constructo y seleccionando las respuestas que aparecen con mayor frecuencia como relevantes para la definición del mismo.

b) *Revisar las investigaciones publicadas:* una revisión de los trabajos realizados con anterioridad acerca del constructo, y de los instrumentos de evaluación existentes, pueden constituir un buen marco de referencia.

c) *Llevar a cabo un análisis de tareas:* ¿qué hacen las personas en sus puestos de trabajo?, ¿qué conductas son las más frecuentes?, ¿qué es lo que distingue a los trabajadores más cualificados de los incompetentes? A partir de estas observaciones se puede elaborar una lista de *conductas críticas* que posteriormente se puede utilizar para la evaluación del rendimiento en el puesto de trabajo.

d) *La observación directa por parte del constructor del test* le permitirá identificar conductas relacionadas con el constructo a medir. Si el constructo a medir fuera la extraversión o la depresión, se podría analizar qué hacen las personas diagnosticadas como deprimidas.

e) *Utilizar el juicio de expertos:* las opiniones de aquellas personas que ya hayan trabajado en el constructo de interés pueden resultar muy valiosas.

f) *Revisar los objetivos instruccionales o de programas de intervención*: la revisión de los contenidos de los programas y/o de los textos u otros materiales permite establecer una serie de objetivos que han de evaluarse en el test y que especifican una serie de conductas que deberán mostrar los sujetos.

Todo esto nos da una idea de la dificultad de definir un constructo y determinar su dominio de conductas. Por otra parte, es fácil deducir que no se trata de dos procesos distintos, sino que están fuertemente relacionados. A medida que se tiene una mayor información acerca del constructo que se quiere medir se irá modificando el contenido de la prueba y viceversa, las investigaciones encaminadas a determinar el dominio de conductas del constructo aumentarán el conocimiento del mismo y permitirán clarificar su concepción.

En los tests de rendimiento y conocimientos y, sobre todo, en los tests referidos al criterio (TRC), en lugar de hablar de un dominio de conductas se hablará de un dominio de contenidos a evaluar y la especificación del mismo cobrará una importancia extraordinaria.

Cuando se quiere construir un test para ser utilizado en el ámbito académico, por ejemplo para la evaluación del rendimiento de los alumnos en una determinada materia, se espera que el test refleje lo que los alumnos han aprendido o lo que deberían haber aprendido a lo largo del curso; en este caso, el camino que se suele seguir para especificar el contenido del test es construir una tabla de doble entrada, en la que las columnas representen las distintas áreas de contenido o facetas del constructo a medir y las filas los distintos procesos u operaciones implicados en la resolución de las tareas planteadas. En algunos casos, como pueden ser pruebas de evaluación, en las filas se puede representar el grado de dificultad de las preguntas que se quieren incluir en la prueba.

Aunque los contenidos de las filas variarán en función de aquello que se quiera medir, se ha intentado hacer una categorización jerárquica, más o menos uniforme, de los distintos procesos implicados. La más popular es la que preparó un comité de la *American Educational Research Association* presidido por Benjamín Bloom (1954), que recoge los siguientes procesos ordenados jerárquicamente en función de su mayor o menor complejidad:

- *Conocimiento*: Recuerdo del material presentado. Por ejemplo: Los ríos de España.
- *Comprensión:* Interpretación o extrapolación de un concepto de forma distinta a la originalmente presentada.
- *Aplicación:* Resolución de problemas nuevos mediante la aplicación de principios familiares o generalizaciones.
- *Análisis:* Descomposición de un problema en sus componentes elementales. Este proceso requiere reconocimiento de elementos múltiples y de las relaciones y/o principios de organización entre los elementos.
- *Síntesis:* Combinar elementos a partir de una estructura original o resolver un problema que requiere la combinación secuencial de varios principios.
- *Evaluación*: Empleo de criterios internos (generados por el sujeto) o externos para hacer juicios críticos.

Esta categorización completa no siempre es utilizable, hay veces que no es necesaria la evaluación de tantos procesos en cuyo caso es posible hacer una selección de aquellos que se considere más importantes o, simplemente, elegir los más adecuados. Sin embargo, la lista completa facilita la preparación y selección de los ítems que van a formar parte de la prueba.

EJEMPLO:

Supongamos que queremos elaborar un test para evaluar el nivel de conocimientos alcanzado por nuestros alumnos en la asignatura de Psicometría. Para ello, elaboramos una tabla de doble entrada (tabla 2.3) situando en las columnas las distintas áreas de contenido de la Psicometría incluidas en los textos utilizados, y en las filas los distintos procesos (se trata de un ejemplo ficticio en el que no se han incluido todas las áreas de la Psicometría ni todos los procesos). En lugar de los procesos se podrían haber utilizado otros criterios para la selección de los ítems, por ejemplo la dificultad de los mismos, tal y como se ha comentado anteriormente: fáciles, de dificultad media o difíciles.

Es importante que estén representadas todas las áreas de la Psicometría incluidas en los textos utilizados para la preparación de la asignatura para asegurar que la prueba que se elabore tenga validez de contenido y que, en cada una de ellas, haya un número suficiente de ítems para asegurar una alta fiabilidad. Sin embargo, es necesario tener en cuenta las limitaciones que pueda marcar un test demasiado largo tanto en lo que se refiere al tiempo, a la motivación de los sujetos, e incluso a las limitaciones debidas a las características de los mismos como puede ser la edad, su salud, etc.

TABLA 2.3
Matriz de especificaciones del contenido para un test de Psicometría

Procesos	Áreas de contenido			
	Medición (20%)	Fiabilidad (35%)	Validez (35%)	A. de ítems (10%)
Conocimiento (25%)	4	7	7	2
Comprensión (25%)	4	7	7	2
Aplicación (25%)	4	7	7	2
Análisis (25%)	4	7	7	2
Total	16	28	28	8

Hay veces que es difícil hacer una categorización completa de las distintas áreas de contenido, y otras veces es difícil establecer los puntos de corte entre un área y otra, de manera que las categorías correspondientes a cada área de contenido sean mutuamente exclusivas. En este caso, es conveniente consultar con expertos para llevar a cabo la categorización.

Cada casilla de la matriz representa la interacción entre un área de contenido y un proceso, esto facilita la redacción de ítems que cubran todos los aspectos que se consideran importantes. No obs-

tante, hay otro punto que hay que aclarar: si a todos los procesos y a todas las áreas de contenido se les concede la misma importancia; si esto es así, será necesario cubrir todas las celdas de la matriz con el mismo número de ítems. Por el contrario, si se considera que no todas las áreas son igualmente importantes, ni todos los procesos, será necesario establecer una ponderación para poder establecer el número de ítems de cada casilla.

Supongamos que en nuestro ejemplo las áreas de fiabilidad y validez son las más importantes, que después es la medición el área más importante y, por último, el análisis de ítems. En este caso una ponderación que se podría dar sería 35% de los ítems para fiabilidad, 35% para validez, 20% para el área de medición y un 10% para el análisis de ítems. Si se asume que todos los procesos tienen la misma importancia para nosotros, a cada uno de ellos le correspondería una ponderación del 25%.

Una vez asignadas las ponderaciones es necesario decidir el número de ítems que va a tener la prueba piloto. Si queremos que tenga 80 ítems (hay que tener en cuenta que en algún tipo de pruebas, la versión piloto debe contar como mínimo con el doble o triple de los ítems de la versión final del test), entonces, en función de los pesos asignados a cada área, hay que establecer el número de ítems que hay que elaborar para cada una de ellas. En nuestro caso sería el 20% para el área de medición (16 ítems), un 35% para fiabilidad y un 35% para validez (28 ítems para cada una de ellas) y un 10% para el análisis de ítems (8 ítems). Para cada uno de los procesos habría que construir 20 ítems puesto que todos tienen la misma ponderación. Ahora bien, de los 16 ítems correspondientes al área de medición, un 25% corresponden a cada uno de los procesos, así habrá que construir 4 ítems por cada uno de los procesos. Esos valores son los que aparecen en el interior de cada casilla de la matriz. Los valores del resto de las casillas se obtendrían de la misma forma.

La especificación del contenido a la hora de construir un test de aptitudes tiende a ser menos específica puesto que está pensada para medir una característica más general y persistente de la persona. A veces lo que se especifica es el tipo de ítems que la van a conformar. Por ejemplo, si se quiere construir un test para medir la *habilidad verbal*, constructo que se había definido anteriormente, se puede especificar que los ítems que lo formen sean de analogías verbales, sinónimos y antónimos, ítems de frases incompletas, etc.

Cuando lo que se quiere construir es un test (escala, cuestionario, inventario) para medir constructos de naturaleza no cognitiva: intereses, actitudes, temperamento, etc., las especificaciones pueden ser bastante esquemáticas y a veces el contenido de los ítems se deduce fácilmente de la misma definición del constructo. La definición ofrecida por Thorndike acerca del constructo *impulsividad*, y que hemos expuesto anteriormente, sugiere que los ítems hagan referencia, entre otros, a los siguientes aspectos: *a)* rapidez en las decisiones, *b)* interés por las actividades no planificadas, *c)* desagrado por las cosas y actividades muy planificadas, etc. En este tipo de tests hay veces que a partir del análisis y reflexión sobre las teorías existentes acerca del constructo que se quiere medir surgen los ítems que se deben incluir.

5.2. Formato de los ítems

Una vez terminada la etapa anterior ya se sabe lo que se va a medir, ahora hay que dar respuesta a la pregunta de cómo medirlo. Para ello, el primer paso es seleccionar el tipo de ítems que se van a utilizar para construir el test teniendo en cuenta que, como señala Osterlind (1998), el ítem puede ser considerado como una unidad de medida dentro del test global y puede aparecer bajo diferentes formatos.

Haladyna (1994), considera que una primera aproximación es establecer dos grandes categorías de ítems en función del tipo de respuesta que exijan de los sujetos: ítems de elección e ítems de construcción.

Los *ítems de elección* son ítems de respuesta cerrada, en ellos se exige a los sujetos que respondan eligiendo una o varias alternativas de entre las propuestas. En los *ítems de construcción*, el sujeto deberá elaborar su propia respuesta. Cada una de estas grandes categorías incluye distintos tipos de formatos, que serán más o menos adecuados dependiendo de si la variable que se quiere medir es de tipo cognitivo o bien de tipo oréctico o afectivo.

5.2.1. Ítems de elección

Los formatos más comunes que presentan son:

- *Dos alternativas:*

Los sujetos han de elegir una entre dos alternativas. Por ejemplo: Verdadero-Falso, Si – No, Correcto – Incorrecto.

EJEMPLO:

— La capital de España es Madrid ..... Si No

— Los constructos son variables observables directamente ..... V F

Este tipo de formato se utiliza normalmente para medir variables de tipo cognitivo: habilidades, aptitudes y, fundamentalmente para la construcción de test de conocimientos y rendimiento. Presenta la ventaja de que es rápido y fácil de usar, pero tiene el inconveniente de que los sujetos que no conozcan la respuesta y respondan al azar tienen un 50% de posibilidad de elegir la respuesta correcta.

No es el formato adecuado para los tests de personalidad y los de actitudes, intereses, etc., es decir, para los tests destinados a medir variables no cognitivas, dado que en ellos no hay respuestas correctas ni incorrectas y, por otra parte, un rango de respuestas restringido a dos alternativas no es el más adecuado para reflejar la posición de los sujetos en este tipo de variables.

- *Elección múltiple*:

Este tipo de ítems consta de : *a)* el enunciado propiamente dicho y *b)* las alternativas u opciones de respuesta, que consisten en una lista de posibles respuestas de las cuales una es la correcta, o la más adecuada, y las otras son incorrectas y se denominan *distractores*. Se suelen utilizar de 3 a 5 alternativas de respuesta para disminuir la posibilidad de que los sujetos elijan la alternativa correcta por azar. Por ejemplo, en un ítem con cinco alternativas de respuesta, de las cuales sólo una es correcta, los sujetos tienen una posibilidad de acertarlo por azar de un 20% (1/5) frente a la del 50% (1/2) que tendrían si el ítem hubiera tenido dos alternativas. También este tipo de formato se utiliza para medir variables cognitivas y fundamentalmente en tests de conocimientos y rendimiento. No se utiliza para medir variables de personalidad, intereses, actitudes, etc., es decir, en el ámbito oréctico o afectivo.

Presentan la ventaja de que son fáciles de administrar, corregir y puntuar y hoy día se pueden corregir mediante el uso de lectoras ópticas y los programas de ordenador adecuados; pero, presentan el inconveniente de que son más difíciles de construir que los de dos alternativas. Es difícil construir alternativas que sean realmente efectivas, hay veces que una alternativa incorrecta es tan obvia que resulta improbable que alguien la elija, con lo cual no está actuando como un distractor. En este caso, si el ítem tiene 5 alternativas puede suceder que, en realidad, esté funcionando como un ítem con 3 ó 4 opciones de respuesta.

Lo ideal sería disponer de un banco de ítems del que se pudieran ir eligiendo, en cada ocasión, los más adecuados.

EJEMPLO:

La capital de España es:

a) Madrid

b) Barcelona

c) La Coruña

d) Sevilla

Muñiz y García Mendoza (2002) muestran una clasificación de los ítems de elección múltiple en función de la estructura que tenga el enunciado y las distintas alternativas.

El enunciado, o base del ítem, puede presentarse en *forma interrogativa, enunciativa* o como una *frase truncada o incompleta*. Cada una de estas formas dará lugar a un tipo de ítem.

La forma más directa y la más recomendable de solicitar una respuesta a los sujetos es la interrogativa. En el estudio que realizan los autores antes citados hacen una revisión de los exámenes PIR de los años 1998, 1999 y 2000 y encuentran que el porcentaje de ítems de forma interrogativa fueron 11%, 10% y 25% respectivamente, un porcentaje muy bajo.

La forma enunciativa es equivalente a la anterior y puede ser utilizada si el conjunto del ítem es coherente; es decir si el enunciado o base del ítem concuerda con las distintas alternativas y éstas son de contenido homogéneo y están bien redactadas.

Los ítems cuya base es una frase incompleta que continúa en alguna de las alternativas que se proponen suele utilizarse en tests educativos puesto que es fácil de construir a partir de frases de los textos.

En relación con la forma de redactar las alternativas, hay dos tipos de ítems: los que presentan una única respuesta correcta y aquellos en los que todas las alternativas son parcialmente correctas pero hay una que es más completa y la mejor respuesta de las presentadas. El primer tipo de ítems se suele utilizar cuando no hay ambigüedad acerca de la veracidad o falsedad de la respuesta, y el segundo cuando se pretenden evaluar procesos mentales complejos.

- *Emparejamiento*:

Este formato implica que el sujeto empareje los elementos de dos columnas de acuerdo a las instrucciones dadas en el enunciado. Al igual que los formatos presentados anteriormente, está indicado para medir variables de tipo cognitivo y, sobre todo, conocimientos.

EJEMPLO:

Seleccione de la columna de la derecha la ciudad española que pertenece a la Comunidad Autónoma situada en la columna de la izquierda y ponga en el espacio en blanco que aparece al lado de cada Comunidad la letra asignada a la ciudad que le corresponde:

1) Castilla-León ..... a) Santander

2) Cantabria ..... b) Segovia

3) Andalucía ..... c) Cáceres

4) Extremadura ..... d) Sevilla

- *Formato Cloze o incompleto:*

En este tipo de ítems se ofrece a los sujetos, por ejemplo, un párrafo o una frase en la que faltan algunas palabras y aparece un espacio en blanco en su lugar, a continuación, se ofrece una lista en la que se incluyen las palabras que faltan. La tarea de los sujetos consiste en seleccionar, de la lista de palabras que se le ofrece, la que corresponda a cada espacio en blanco.

EJEMPLO:

En el río había gran cantidad de ..... que navegaban en ambas direcciones. No se podía estar en cubierta debido al fuerte ....., pero como el trayecto no era muy ..... no era demasiado molesto permanecer en el/la .....

a) Barcos
b) Interior
c) Viento
d) Largo

Nota: Recordar que los formatos incluidos hasta ahora se utilizan, fundamentalmente, para la medida de habilidades, aptitudes y conocimientos. En ellos, se decide de antemano cuál es la respuesta correcta y las incorrectas. Los tests elaborados con este tipo de ítems se denominan tests objetivos.

• *Escalas de clasificación (rating scales):*

Se trata de un tipo de formato de ítems en el que se presenta un enunciado y distintas alternativas de respuesta que están ordenadas de forma gradual en una serie de categorías a lo largo de un continuo. El sujeto debe responder eligiendo, de entre las alternativas propuestas, aquella que mejor refleje su postura o actitud personal ante el enunciado.

Dado que a los sujetos se les pide que emitan juicios de valor, puesto que han de mostrar su postura personal, a este tipo de escalas se las denomina *escalas valorativas*.

EJEMPLO:

El tabaco debería prohibirse en todos los sitios públicos:

a) Totalmente de acuerdo
b) De acuerdo
c) Me es indiferente
d) En desacuerdo
e) Totalmente en desacuerdo

Aunque este formato se parece al que presentan los ítems de elección múltiple, en cuanto que hay un enunciado y varias opciones de respuesta, hay una diferencia muy clara entre ellos. En los ítems de elección múltiple las distintas opciones son independientes entre sí; por el contrario, las opciones de las escalas de clasificación son interdependientes y corresponden a categorías de respuesta ordenadas gradualmente.

Este tipo de formato no se utiliza en el ámbito cognitivo, ya que no implica respuestas correctas o incorrectas, sino para medir variables no cognitivas: actitudes, intereses, personalidad, etc.

Tienen la ventaja de que los sujetos expresan su postura de una manera más precisa que en los ítems de elección múltiple; pero tienen también sus inconvenientes; uno de ellos, muy importante,

es que el significado de las distintas opciones de respuesta no es el mismo para todos los sujetos. Por ejemplo, la alternativa *de acuerdo* no siempre significa lo mismo para todos. Por otra parte, es frecuente que aparezcan *sesgos* en las respuestas; es decir, hay sujetos que tienden siempre a elegir las opciones extremas o, por el contrario, cuando hay un número impar de categorías algunos sujetos tienden a elegir la categoría central.

Respecto al número de opciones más adecuado no hay un acuerdo generalizado, pero lo cierto es que cuando hay más de 7 los sujetos se sienten incapaces de diferenciar entre los significados de las categorías contiguas. En general, el tipo de formato más utilizado es el de 5 alternativas de respuesta propuesto por Likert en 1929 para la elaboración de escalas de actitudes. Osgood (1952,1976) en su técnica denominada Diferencial Semántico utilizó 7 categorías de respuesta.

Las etiquetas lingüísticas asignadas a las distintas categorías variarán dependiendo del tipo de escala utilizada, en general reflejan los siguientes aspectos:

Acuerdo: Totalmente en desacuerdo ..... Totalmente de acuerdo
Frecuencia: Siempre ..... Nunca
Cantidad: Mucho ..... Nada
Sentimientos: Completamente satisfecho ..... Completamente insatisfecho
Valoración: Excelente ....., Muy mala

Entre medias de esas categorías extremas se irán asignando distintas etiquetas lingüísticas en función del número de alternativas.

• *Listados (checklists):*

Se trata también de una escala valorativa en la que los sujetos han de mostrar su opinión respecto a algún hecho (idea, objeto, persona, etc.) presentado en el enunciado. No se utilizan para la medida de variables de tipo cognitivo. A diferencia de las escalas de clasificación, las opciones de los listados no están ordenadas sino que son independientes entre sí. También se diferencian de los ítems de elección múltiple en que en los listados no hay respuestas correctas o incorrectas.

Por otra parte el número de alternativas de respuesta suele ser bastante grande (una lista) y no siempre es necesario elegir una única opción, es posible elegir varias opciones. Es un formato típico de los cuestionarios.

EJEMPLOS:

— En su opinión, cuál de los deportes que aparecen a continuación es su preferido:
  a) Natación
  b) Fútbol

c) Tenis

d) Golf

Los sujetos deberán responder marcando la alternativa elegida.

— De los adjetivos que aparecen a continuación, señale con una cruz aquellos que mejor le definan:

a) Simpático	e) Sociable
b) Tímido	f) Estudioso
c) Paciente	g) Trabajador
d) Impulsivo	h) Perezoso

Nota: Las escalas de clasificación y los listados se utilizan para la medida de variables de personalidad, actitudes, opiniones, etc. Variables no cognitivas. En este tipo de pruebas no hay respuestas correctas o incorrectas.

5.2.2. Ítems de construcción

En este tipo de ítems es el propio sujeto el que ha de elaborar su respuesta, de ahí que se denominen de respuesta abierta. Ahora bien, dentro de esta categoría de ítems hemos de distinguir los de respuesta corta y los de respuesta extensa o de ensayo.

- *Ítems de respuesta corta*:

A veces no son más que modificaciones de los ítems de elección múltiple pues el sujeto ha de elegir una única palabra; pero, en lugar de elegirla de entre una serie de alternativas que se le ofrecen, la tiene que construir él mismo; otras veces el sujeto tiene que responder con una frase.

EJEMPLO:

— El nombre del presidente de Gobierno español es.....

- *Ítems de respuesta extensa o de ensayo:*

Se pide a los sujetos, por ejemplo, que desarrollen un tema.

EJEMPLO:

Describa el origen y desarrollo de la Teoría de los Tests.

Dado que la realidad de la vida es algo bastante complicado, no siempre es adecuado utilizar un formato de respuesta cerrada en los ítems pues la información que ofrecen es una información parcial (Makel, 1998). A veces es preferible dar a los sujetos la oportunidad de que expresen con sus propias palabras sus conocimientos, experiencias, opiniones, etc. y, de esta manera, el investigador podrá conocer no sólo lo que saben, piensan y opinan acerca de aquello sobre lo que se les pregunta, sino cómo lo expresan, pudiendo también analizar ciertos aspectos de la respuesta como puede ser la originalidad, la forma de redactar, etc., que en determinadas situaciones son cualidades necesarias. Es decir, se podrá evaluar no sólo el nivel de conocimientos de los sujetos y su forma de estructurarlos, sino sus habilidades cognitivas de orden superior, los procesos cognitivos que ponen en marcha a la hora de solucionar un problema.

Este tipo de formato de los ítems se utiliza para medir todo tipo de variables, tanto cognitivas como orécticas y afectivas, pero tiene un inconveniente importante y es que las respuestas de los sujetos son más difíciles de analizar y valorar que las de los ítems de respuesta cerrada, puesto que el investigador tiene que codificarlas en una serie de categorías antes de comenzar el análisis. La codificación incluye agrupar juntos a los sujetos que han emitido respuestas similares y es muy difícil encontrar a dos sujetos que hayan dado la misma respuesta. En este caso el investigador suele emitir juicios subjetivos acerca de lo que los sujetos querían o no decir cuando emitieron sus respuestas.

En cuanto a la dificultad de construcción a nadie se le escapa que es mucho más fácil preparar este tipo de pruebas que los tests objetivos, de ahí que cuando la población a la que se dirige el test es pequeña se suelen utilizar tests con ítems de respuesta corta (Nunnally y Bernstein, 1995).

5.3. Longitud del test

Al hacer la matriz de especificaciones del contenido (ver tabla 2.3) se explicó la forma en que se podía calcular y repartir el número de ítems de un test, partiendo de un número inicial de ítems, en función de las áreas de contenido, de los procesos que se iban a evaluar o de cualquier otra variable que se quiera tener en cuenta a la hora de construir un test. Ahora bien, ¿cuál es el número de ítems adecuado en cada caso? Realmente no hay una respuesta única a esta pregunta, ya que son muchos los factores que hay que tener en cuenta: la población a la que va dirigido, las limitaciones de tiempo, los objetivos del test, etc.

En cuanto a la *población a la que va dirigido* no es lo mismo construir un test para ser utilizado en una población infantil que en una población adulta, no sólo el tiempo que tardan los niños en procesar la respuesta a cada ítem y en escribirla es distinto, sino que también varía su capacidad de atención y motivación. Es muy difícil conseguir que los niños puedan responder correctamente a tests muy largos.

El tiempo del que se dispone también es otro factor a tener en cuenta a la hora de fijar la longitud del test. Si se asume que los bachilleres o universitarios tardan aproximadamente 1 minuto en responder a un ítem de elección múltiple en un test de conocimientos, difícilmente se podrá poner un test de más de 60 ítems cuando se cuente con menos de una hora de tiempo para su realización. Como norma general se debería asumir que, a no ser que lo que se desee medir sea la rapidez de respuesta de los sujetos, la longitud del test debe ser tal que todos tengan tiempo suficiente para intentar resolver o contestar a todos los ítems.

Los objetivos del test es otro factor a tener en cuenta, si el test se quiere construir para medir un área de conocimiento muy concreta deberá estar formado por ítems muy específicos y similares, pero no será necesario que sea muy largo; sin embargo, si el test debe cubrir varias áreas de contenido deberá incluir un mayor y más variado número de ítems.

La matriz de especificaciones del contenido nos puede dar una idea acerca del número de ítems a incluir. En lugar de partir del número de ítems que debe tener la prueba piloto para hacer el reparto de ítems en cada casilla, como se hizo anteriormente, se puede proceder a la inversa; se puede partir del número mínimo de ítems que ha de tener una de las casillas y, teniendo en cuenta los factores de ponderación asignados a cada área de conducta y a cada proceso a evaluar, se van calculando el número de ítems del resto de las casillas. Al final se podrá contar con el número de ítems del test total.

En cualquier caso, se recomienda que en la prueba piloto se incluya un número de ítems que sea mayor que el que se va a utilizar en la versión final, ya que a lo largo de los distintos análisis que se deberán ir haciendo se irán eliminando aquellos ítems que no reúnan las propiedades psicométricas adecuadas.

5.4. Características psicométricas de los ítems

Cuando hablamos de características psicométricas de los ítems nos referimos fundamentalmente a su nivel de dificultad, a su homogeneidad en relación con los demás ítems que formen el test y a su capacidad de discriminación. Aunque no vamos a entrar en la explicación de los métodos estadísticos que implica su cálculo, puesto que son aspectos que se irán analizando en los temas siguientes, sí queremos hacer referencia a su significación y a la importancia que tienen a la hora de seleccionar los ítems para la construcción de un test.

En el marco de la Teoría Clásica de los Tests, diremos que un ítem es fácil o difícil para una determinada población, en función de la probabilidad que tengan los sujetos de responder a él correctamente. Si esta probabilidad es alta, el ítem será fácil y, por el contrario, será difícil si la probabilidad es baja.

Un ítem tendrá un alto grado de homogeneidad con el resto de los ítems que formen el test cuando mida lo mismo que ellos.

Un ítem tendrá poder discriminativo en la medida en que sirva para diferenciar entre sujetos que han obtenido en el test puntuaciones extremas.

Aunque estas características se analizarán en profundidad en un tema posterior, han de tenerse en cuenta a la hora de construir un test, pues dependiendo del uso que se le vaya a dar será necesario que los ítems seleccionados tengan unas características determinadas.

Respecto a la dificultad de los ítems vamos a hacer una distinción entre tres tipos de tests: de velocidad, de ejecución máxima y de ejecución típica.

Tests de velocidad: En este tipo de tests los ítems deben ser muy fáciles de resolver, la dificultad estriba en que tienen un tiempo limitado de ejecución y este es el factor que va a permitir diferenciar y discriminar entre los sujetos. Si no existiera limitación del tiempo, la mayoría de los sujetos serían capaces de resolver correctamente todos los ítems. Algunos tests construidos para medir variables cognitivas son tests de velocidad; por ejemplo un test que mida rapidez de cálculo.

Tests de ejecución máxima (Tests de potencia). Utilizados fundamentalmente para la evaluación del rendimiento académico y para la medida de las aptitudes y destrezas. En este tipo de tests los ítems presentan diferentes grados de dificultad, desde ítems muy fáciles que puedan ser respondidos por todos los sujetos y que deberán estar situados al comienzo de la prueba, hasta ítems muy difíciles que no puedan ser acertados más que por los sujetos más aptos y que se colocan al final de la prueba. En este tipo de tests el tiempo no es un factor que deba influir. Los sujetos han de tener el tiempo suficiente para poder intentar resolver todos los ítems, y si no lo hacen no debe ser por falta de tiempo sino porque no conocen la respuesta.

Tests de ejecución típica: Son los tests de personalidad, actitudes, intereses, etc. Dado que en ellos no hay respuestas correctas o incorrectas no tiene sentido hablar de dificultad de los ítems.

El grado de homogeneidad de los ítems depende del constructo que se quiera medir con el test. Si se trata de un constructo unidimensional los ítems han de ser más homogéneos que si el constructo a medir es multidimensional. Si el constructo es multidimensional y todos los ítems del tests miden una única dimensión, habrá aspectos del mismo que no serán medidos y, por lo tanto, las inferencias que se hagan a partir de las puntuaciones que obtengan los sujetos en el test no serán lo suficientemente válidas.

En cuanto al nivel de discriminación de los ítems dependerá de la población a la que va dirigido el test. Si el testestá dirigido a la población general será necesario que los ítems permitan discriminar entre los distintos niveles (de rendimiento, conocimientos, aptitud o destreza) que presenten los sujetos. Esto quiere decir que el test deberá estar formado por ítems que cubran todos los niveles de dificultad, desde los más fáciles a los más difíciles. Dado que los ítems que más discriminan en este tipo de poblaciones son los de dificultad media, el mayor porcentaje de ítems deberá tener este grado de dificultad.

Si se quiere que el test detecte a los sujetos más brillantes y discrimine entre ellos, los ítems deberán ser difíciles y muy difíciles, de manera que los sujetos que presenten un nivel medio y bajo no los puedan responder correctamente y sólo lo hagan los más capacitados.

Si, por el contrario, ahora se quisiera discriminar entre los menos capacitados, los ítems deberían ser fáciles y muy fáciles, de manera que sólo los fallaran los menos capacitados.

6. REDACCIÓN DE LOS ÍTEMS

La realidad es que si queremos construir un buen test hay que tener en cuenta una cosa, que si los ítems que lo van a formar son malos el test no puede ser bueno, de ahí la importancia de cuidar la redacción de los mismos.

Algunos autores piensan que la construcción de ítems es un arte que pocas personas dominan (Nunnally y Bernstein, 1995); sin embargo, hay una serie de consideraciones que pueden ayudarnos en la tarea:

1. Debe existir un alto grado de congruencia entre el ítem y el constructo psicológico que se quiere medir (validez de constructo).
2. Los constructos deben estar claramente definidos. Si no es así difícilmente se podrá valorar el grado de congruencia ítem-constructo.
3. Hay que tratar de minimizar los errores de medida cometidos al medir el constructo con cada ítem.
4. El formato de los ítems ha de ser adecuado para los objetivos del test.
5. Los ítems deben reunir las características psicométricas más adecuadas en cada caso.
6. Los ítems deben estar bien redactados.
7. Los ítems deben satisfacer las consideraciones legales y técnicas pertinentes. Por ejemplo se deben evitar los plagios.

Las cinco primeras consideraciones se han abordado ya, en cierto modo, a lo largo del tema; ahora vamos a ocuparnos de la redacción de los ítems.

Todas las personas que nos hemos enfrentado a la tarea de escribir, nos damos cuenta de la dificultad que entraña, y del número de veces que hay que rehacer el texto hasta que estamos más o menos conformes con lo escrito. Si se trata de escribir algo técnico, por ejemplo redactar los ítems de un test, la dificultad todavía es mayor porque se requiere un alto grado de precisión en el uso del lenguaje (Osterlind, 1998). Ahora bien, para poder alcanzar ese grado de precisión es necesario tener un conocimiento profundo del contenido al que van a hacer referencia los ítems. Yo

podré ser una «artista» escribiendo, pero desde luego no podría construir buenos ítems para medir el conocimiento de los alumnos de Ingeniería Industrial en la asignatura de Resistencia de Materiales porque mi desconocimiento del tema es absoluto.

Una vez que se presupone el conocimiento del contenido que han de tener los ítems del test, para poder redactar buenos ítems conviene aceptar una serie de recomendaciones, unas de carácter general y otras específicas del tipo de formato que se vaya a utilizar en la redacción.

6.1. Recomendaciones generales

Aunque algunas pueden parecer obvias y de sentido común la experiencia demuestra la necesidad de recordarlas.

- *Evitar la ambigüedad de los enunciados*

Una forma de hacerlo es redactándolos de forma clara. El significado de las palabras empleadas debe estar claro para todos los sujetos ya que difícilmente serían comparables sus respuestas si cada uno pudiera interpretar de manera distinta el significado del enunciado. Términos como *religiosidad* o *patriotismo,* por ejemplo, pueden ser interpretados de manera diferente por distintos sujetos; entonces, cuando se aluda a ellos tiene que quedar muy claro a qué se está haciendo referencia (Fowler, 1995; Weisberg et al., 1996).

Los enunciados cortos y directos también contribuyen a evitar la ambigüedad ya que la inclusión de palabras innecesarias complican la lectura y pueden provocar confusión en los sujetos (Payne, 1951).

Es necesario que sean lo más precisos posible. Hay ítems que incluyen preguntas acerca de las actividades de las personas en los *últimos años*, o sus proyectos para los *próximos años*. Esto provoca ambigüedad y es necesario precisar qué se entiende por los últimos años o los próximos años. Esos términos pueden tener distinto significado para los sujetos ya que mientras para unos pueden significar 2 años, para otros pueden ser 5 o 10 años.

- *Evitar enunciados que provoquen respuestas sesgadas*

Se deben evitar los enunciados que puedan provocar una respuesta sesgada, entendiendo por respuesta sesgada aquella que es más probable que elijan los sujetos independientemente de su opinión. Por ejemplo un enunciado que implique que los sujetos deban admitir conductas o actitudes que no son consideradas socialmente aceptables puede provocar que los sujetos no manifiesten su verdadera opinión y elijan la respuesta socialmente aceptable.

• *Expresar una única idea en el enunciado*

Es necesario evitar las dobles preguntas en un mismo enunciado ya que provocaría confusión en los sujetos y no sabrían qué respuesta emitir.

EJEMPLO:

Está usted a favor de reducir el consumo de alcohol entre los jóvenes y aumentar los impuestos de las bebidas alcohólicas ..... SÍ NO

Este enunciado es incorrecto, se incluyen dos conceptos diferentes. Una persona puede estar a favor de reducir el consumo de alcohol entre los jóvenes pero no a base de aumentar los impuestos, con lo cual no sabría que opción elegir. De un único enunciado se podrían obtener dos ítems:

a) Está usted a favor de reducir el consumo de alcohol entre los jóvenes .... SÍ NO

b) Está usted de acuerdo en que se aumenten los impuestos de las bebidas alcohólicas para reducir su consumo entre los jóvenes..... SÍ NO

• *Evitar las dobles negaciones en los enunciados:*

En general es preferible no abusar de los enunciados negativos, pero lo que sí que hay que evitar es el uso de las dobles negaciones ya que provocan que los sujetos no sepan cual es la respuesta que representa su punto de vista u opinión.

EJEMPLO:

Le parece a usted posible o imposible que la llegada del hombre a la luna nunca hubiera ocurrido ..... V F

Ante este enunciado uno no sabría que responder, sería imposible.

6.2. Recomendaciones para ítems de elección

Además de las normas generales anteriormente expuestas hay una serie de normas específicas para cada tipo de formato.

• *Dos alternativas: Verdadero-Falso*

1. Estar absolutamente convencido de que el ítem es sin ninguna duda verdadero o falso.

EJEMPLO:

Dalí fue el mejor pintor del siglo veinte ..... V F

Se trata de un enunciado mal elaborado ya que eso es muy subjetivo.

2. No utilizar frases que sean universalmente verdaderas o falsas.
3. Evitar en el enunciado palabras que puedan, de alguna manera, inducir la respuesta correcta a los sujetos aunque no la conozcan.

Términos como *siempre, todo, nada, nunca, exclusivamente,* inducen la respuesta ya que suele ocurrir que cuando se utilizan estos términos en un ítem de dos alternativas (Verdadero-Falso) el ítem es falso. Por el contrario términos como *a veces, en general, apenas....* hacen mucho más probable que el enunciado del ítem sea verdadero.

4. Situar a lo largo del test, de forma aleatoria, los ítems cuyo enunciado sea correcto; de esta manera se evitan patrones de respuesta reconocibles por los sujetos. Por ejemplo, si se introdujera un ítem falso cada dos ítems verdaderos y los sujetos descubren la secuencia del patrón pueden responder correctamente a un ítem sin conocer la respuesta.

• *Elección múltiple*

1. Asegurarse de que el enunciado del ítem formula el problema con claridad.
2. Incluir la mayor parte del texto en el enunciado para evitar repeticiones innecesarias en las opciones de respuesta.
3. Incluir las distintas opciones de respuesta al final del enunciado.
4. Asegurarse de que los distractores (alternativas incorrectas) son plausibles.
5. Evitar opciones de respuesta como *Ninguna de las anteriores, Todas las anteriores.*
6. Que sólo haya una opción correcta (o más correcta), a no ser que se indique lo contrario claramente en las instrucciones.
7. Tratar de que todas las alternativas de respuesta tengan una longitud aproximadamente igual y con una construcción gramatical parecida.
8. Aleatorizar la ubicación de la alternativa correcta.
9. Hacer que todas las alternativas le parezcan igualmente atractivas a una persona no informada del problema al que alude el enunciado.
10. Asegurarse de que cada alternativa concuerda gramaticalmente con el enunciado del ítem. Si el enunciado está en singular, asegurarse que cada alternativa está en singular.

• *Emparejamiento*

1. Asegurarse que tanto las premisas como las opciones de respuesta que hay que emparejar son homogéneas.

   Supongamos que hay dos premisas que hacen referencia a una fecha y en las opciones de respuesta sólo hay dos que incluyen los años; lógicamente el problema se reduce a empa-

rejar esas dos premisas con las dos opciones de respuesta y no hace falta examinar ninguna de las demás.

EJEMPLO:

Premisas	*Opciones*
1. España	a. Berlín
2. Fecha del descubrimiento de América	b. París
3. Francia	c. 1492
4. Alemania	d. Madrid

Aunque un poco exagerado el ejemplo, dado que no hay más que una fecha se sabe con qué premisa hay que emparejar esa opción.

2. Utilizar el formato adecuado.

   Las premisas se deben presentar de forma aleatoria en una columna a la izquierda y en una columna paralela, situada a la derecha, se deben presentar las distintas alternativas de respuesta. Para facilitar la tarea del sujeto se debe dejar un espacio en blando detrás de cada premisa numerada para poder situar la letra correspondiente a la alternativa de respuesta.

3. El enunciado del ítem debe reflejar claramente la tarea que se espera del sujeto y la forma en que hay que llevar a cabo el emparejamiento.

- *Formato Cloze o incompleto*

Es necesario que en el enunciado del ítem haya tantos espacios en blanco como alternativas de respuesta, y en caso de que esto no suceda habrá que hacerlo constar en las instrucciones.

- *Escalas de Clasificación*

1. Evitar expresiones coloquiales en los enunciados de los ítems pues puede haber sectores de la población que las desconozcan.
2. Incluir en el test completo aproximadamente el mismo número de ítems formulados de manera positiva y negativa.

   Dado que las escalas de clasificación se utilizan, fundamentalmente, para la medida de actitudes, opiniones, valores, etc., el test deberá incluir el mismo número de ítems que denoten una actitud positiva o favorable a lo que se está evaluando y una actitud contraria; evitando, como ya se ha comentado anteriormente, las negaciones en el enunciado. Por ejemplo en lugar de poner: *No me gusta mucho la caza* que resulta ambigua, quedaría mejor redactado si pusiera *Odio la caza* con las distintas etiquetas lingüísticas asociadas a las distintas categorías de respuesta.

3. Asignar las etiquetas lingüísticas.

   Aunque ya se ha abordado este tema es importante recordarlo. Teniendo en cuenta que las categorías están ordenadas, hay veces que sólo se incluyen valores numéricos. Es importante que al menos en los extremos de la escala aparezcan las etiquetas lingüísticas pues facilitan la respuesta de los sujetos. También es conveniente introducir una categoría central que represente el punto medio o neutral (por ejemplo *No se, Indiferente, Ni de acuerdo ni en desacuerdo)* pues refleja la opinión o actitud de muchas personas.

- *Listados*

Son fáciles de construir y su redacción no reviste ningún problema. Es necesario seguir las recomendaciones generales.

6.3. Recomendaciones para los ítems de construcción

- *Ítems de respuesta corta*

1. Asegurarse de que el enunciado del ítem puede ser contestado con una única frase o palabra y que hay una única respuesta correcta. Omitir sólo palabras clave.
2. Los espacios en blanco para las respuestas han de ser de la misma longitud. La corrección se facilita si estos espacios se presentan en una columna a la derecha de los enunciados.
3. Evitar dar pistas o claves acerca de la respuesta correcta. Si por ejemplo la respuesta correcta lleva un artículo delante, en el enunciado deberá aparecer así: *el (la), un (una)*, para evitar que los sujetos al responder tengan una pista acerca de la palabra que deben elegir.
4. Indicar el grado de precisión exigido en la respuesta. Si, por ejemplo, la respuesta al ítem requiere hacer cálculos numéricos con decimales, es necesario expresar el número de decimales que se deben utilizar.
5. Evitar determinantes específicos como *Todo o Nada* y ambiguos como *Frecuentemente o Algunas veces.*

- *Ítems de respuesta extensa o de ensayo*

1. Asegurarse de que el problema está bien enfocado. Se debe comenzar el enunciado con palabras que definan claramente la tarea, por ejemplo: *Compare, Contraste....*

   Los sujetos han de saber perfectamente que es lo que se les está preguntando, de esta manera se evitaran las vaguedades en las respuestas.

   En los tests de rendimiento y conocimientos, bajo la presión de una situación de examen, los estudiantes trabajan contra reloj y si no tienen bien delimitado el tema sobre el que deben

hablar es posible que las respuestas sean vagas y pobres. Por otra parte, a medida que está menos estructurada la pregunta es más difícil ser objetivo a la hora de corregirla puesto que la variabilidad de las respuestas es mayor.

2. No permitir a los sujetos que elijan entre varias preguntas de ensayo.

   Si se quiere comparar el rendimiento de los sujetos es necesario hacerlo sobre una tarea común. Si cada sujeto ha tenido opción de elegir responder a ítems diferentes la comparación no es posible.

3. Decidir de antemano cómo se van a puntuar las preguntas de ensayo.

   Este es uno de los grandes problemas de este tipo de ítems ya que es muy difícil conseguir objetividad a la hora de su corrección. Si una prueba de ensayo es corregida por dos personas distintas es fácil que la puntuación asignada varíe considerablemente si no hay unas reglas completas y explícitas acerca de cómo hacerlo.

4. Redactar las preguntas referidas a cuestiones controvertidas de manera que los sujetos que deben responder sean evaluados en relación a la evidencia que presentan no a su posición personal respecto al tema.

6.4. Los sesgos de respuesta

Otro de los aspectos que hay que tener en cuenta a la hora de redactar los ítems, sea cual sea su formato, es la posibilidad de respuestas sesgadas. Es cierto que este tipo de respuestas suelen aparecer en tests construidos para la medida de aspectos orécticos y afectivos: personalidad, intereses, actitudes, etc. Algunos de estos sesgos ya han sido puestos de manifiesto a lo largo del tema pero no está de más recordarlos.

Los principales sesgos de respuestas, que hay que tratar de evitar en lo posible, son producidos por:

— *Aquiescencia* o tendencia a responder sistemáticamente que se está de acuerdo (o en desacuerdo) con el enunciado del ítem con independencia de su contenido.

— *Deseabilidad social* o tendencia a responder al ítem de una manera socialmente aceptable y no en función de lo que uno sienta, opine o piense.

— *Indecisión* o tendencia a seleccionar la alternativa central o neutra correspondiente a etiquetas como *No sé, Ni de acuerdo ni en desacuerdo, Indiferente*. Aunque a veces no es deseable, cuando se observa que la alternativa central puede provocar sesgos de respuesta se puede eliminar.

— *Respuesta extrema* o tendencia a elegir como respuesta las categorías de los extremos con independencia del contenido del ítem.

7. REVISIÓN CRÍTICA POR UN GRUPO DE EXPERTOS

Una vez que se han redactado los ítems del test, y antes de dar forma a la prueba piloto, es conveniente que esos ítems sean revisados por un grupo de personas que no hayan intervenido en su elaboración con el fin de que puedan revisar, no sólo si se adaptan al contenido, sino la claridad de la redacción, si se han cumplido las normas generales y específicas en función del tipo de formato, la corrección de la respuesta *correcta* en los ítems de elección múltiple, la calidad de los distractores elegidos, etc.; en fin, para que analicen todos aquellos aspectos que contribuyen a la calidad del ítem.

Cuando un profesor está implicado en una tarea docente, como puede ser escribir un libro de texto de su asignatura, necesita recibir información acerca de si lo que ha escrito puede ser comprendido por aquellos a los que va dirigido. Puede tener eso que llamamos *deformación profesional* y, debido a su familiaridad con el tema, no ser consciente de que para los demás las cosas no están tan claras. Lo mismo ocurre cuando se quiere confeccionar un examen para evaluar los conocimientos de los alumnos, o cuando se quiere construir otro tipo de pruebas. Pues bien, en todos los casos es conveniente que haya una revisión no sólo en cuanto a los contenidos, sino a su estilo de redacción, dificultad, etc.

Lo ideal es que la revisión pudiera ser hecha por personas expertas, tanto en los contenidos como en estilo de redacción, etc.; si esto no es posible, siempre se podrá contar con alguna persona más o menos cualificada. Si tampoco esto fuera posible, lo mejor es que el constructor haga una segunda lectura de lo escrito al cabo de un cierto tiempo, esta lectura le ofrecerá una nueva visión de su trabajo y le permitirá corregir los posibles errores.

Una vez revisados los ítems y eliminados (o corregidos) aquellos que no fueran considerados idóneos, se puede construir la versión preliminar del test, la prueba piloto, con aquellos que han pasado este primer control de calidad.

8. CONFECCIÓN DE LA PRUEBA PILOTO

Hay algunos aspectos básicos a tener en cuenta para la confección de la prueba piloto: *a)* las instrucciones de administración, *b)* el formato de presentación y de registro de las respuestas.

8.1. Las instrucciones de administración

Salvo raras excepciones, el constructor del test quiere que todas aquellas personas a las que va a ser aplicado entiendan perfectamente lo que deben hacer y que estén motivados para hacerlo; por

eso, a la hora de redactar las instrucciones para la cumplimentación del test se deben tener en cuenta estos objetivos. Cada tipo de pruebas requerirá unas determinadas instrucciones, pero hay algunas que suelen ser bastante comunes y que hemos adaptado de Torndike (1989).

1. Como norma general, a la hora de redactar las instrucciones se debe evitar utilizar lenguajes ampulosos y amenazantes. No se deberá decir por ejemplo: *Esta prueba nos va a permitir conocer lo inteligente que es usted*.
2. En los tests de ejecución máxima, por ejemplo en las pruebas de aptitudes, se debe explicitar que los ítems son de dificultad variable, que hay algunos que resultarán muy difíciles para todos los sujetos, y que la prueba está pensada para que haya ejercicios que no puedan resolver. Si se incluye esta información en las instrucciones se reducirá la ansiedad de los sujetos cuando se enfrenten a este tipo de ítems.
3. En los tests de velocidad, en los que el tiempo está limitado de manera que sólo muy pocos lleguen a completar la prueba, se deberá explicitar también en las instrucciones.
4. Las instrucciones deben proporcionar uno o más ítems como ejemplo, para informar a los sujetos acerca de cómo deben resolver cada uno de ellos y la forma de elegir la solución correcta en caso de que la hubiera. A veces se incluyen también algunos ítems de práctica, sobre todo si se presume que la población a la que va dirigido el test no está familiarizada con este tipo de pruebas.
5. Las instrucciones deben informar acerca de cómo distribuir el tiempo y qué hacer cuando no se conoce la respuesta a un ítem. Cuando hay tiempo límite para responder a la prueba se debe informar a los sujetos para que trabajen con rapidez; no obstante, en cualquier caso no está de más advertirles que no desperdicien mucho tiempo intentando contestar a un ítem cuya respuesta desconocen, que es mejor pasar al siguiente, y que una vez terminada la prueba, si es posible, vuelvan a intentar resolverlos.
6. Las instrucciones deben animar a los sujetos a responder a todas las preguntas y favorecer así su rendimiento, dado que la puntuación de los sujetos tiende a bajar considerablemente cuando se dejan muchas respuestas en blanco. En los ítems de elección múltiple se puede sugerir a los sujetos una doble estrategia, en primer lugar se les puede decir que traten de encontrar la alternativa correcta y, en caso de que no la puedan identificar, que traten de identificar una o más alternativas erróneas, eliminarlas, y analizar las alternativas restantes seleccionando una de ellas.
7. Dado que muchas pruebas se corrigen hoy día mediante hoja de lectora óptica, o se aplican a través del ordenador, las instrucciones deben explicitar claramente la forma de responder en ellas.

8.2. Formato de presentación y de registro de las respuestas

Una vez elaboradas las instrucciones hay que organizar y ordenar los ítems seleccionados para su posterior presentación a los sujetos y decidir el formato de registro de las respuestas. Esta fase que puede parecer trivial también requiere una serie de cuidados.

La forma de registrar las respuestas de los sujetos va a influir, no sólo en las instrucciones, como hemos apuntado anteriormente, sino en el formato final del test. Se puede optar porque los sujetos respondan en la misma hoja o cuadernillo del test o, por el contrario, se puede optar porque respondan en una hoja aparte que les será entregada junto con la hoja o cuadernillo del test. La ventaja de esta última forma de registrar las respuestas es que los tests pueden ser reutilizados. Además, la hoja de respuestas puede ser una hoja de lectora óptica que facilita la corrección de la prueba. En los tests informatizados el registro de las respuestas se hace a través del ordenador.

Si antes hemos dicho que las instrucciones deberían animar a los sujetos a responder a los ítems, la presentación de éstos dentro de la prueba ha de tener el mismo objetivo.

1. El formato de presentación debe ser claro y perfectamente legible por todos los sujetos, evitando que se puedan cometer errores involuntarios como por ejemplo saltarse una pregunta, confundir la casilla de respuesta, etc.
2. Se deben solicitar al comienzo de la prueba los datos de identificación de las personas, su nombre, apellidos, datos de contacto, etc. o, en caso de aplicaciones en las que se requiere el anonimato de los que responden, una clave de identificación.
3. A continuación se presentan las instrucciones para la realización de la prueba, siguiendo las pautas establecidas en el punto anterior.
4. Después de las instrucciones se presentan los ítems:

   En las pruebas diseñadas para medir variables cognitivas (conocimientos, aptitudes, destrezas...) es importante que los ítems estén ordenados en función de su nivel de dificultad. Si al principio de la prueba se pusieran ítems difíciles es posible que muchas personas se sintieran desmotivadas y dejaran de responder.

   En las pruebas diseñadas para medir variables no cognitivas, en las que a veces se incluyen preguntas que pueden resultar embarazosas, es necesario cuidar que éstas no aparezcan al principio de la prueba ya que los sujetos pueden darla por terminada nada más empezar al negarse a contestar.
5. Cuando un mismo test incluye ítems de varios formatos conviene que aparezcan agrupados los de un mismo formato para evitar provocar desconcierto en los sujetos.
6. Hay que tratar de que los ítems sigan una ordenación lógica. Las preguntas referidas a un mismo tema deben situarse unas a continuación de otras de manera que los sujetos no tengan que ir saltando de un tema a otro.

9. APLICACIÓN DE LA PRUEBA PILOTO

Una vez construida la prueba es necesario hacer un estudio piloto de la misma para su evaluación psicométrica; es decir, para ver si cumple los requisitos necesarios que permitan considerarla como un instrumento científico de medición.

La aplicación de la prueba piloto requiere, en primer lugar, decidir acerca de la forma de administración y, en segundo lugar, seleccionar una muestra de sujetos que pertenezcan a la misma población que aquellos para los cuales se ha diseñado el test.

Respecto a la forma de administración de la prueba hay varias posibilidades:

1. *Colectiva - individual*

   Siempre que se pueda hay que tender a que la aplicación pueda hacerse de forma colectiva; no obstante hay algunos tests para adultos y niños que requieren aplicación individual (WAIS y WISC) y algunos manipulativos como *La escala de Alexandre*.

2. *Oral*

   Tanto las instrucciones dadas por el entrevistador como las respuestas emitidas por los sujetos son orales. Se puede hacer de forma personal o bien por teléfono. En el primer caso hay un contacto personal entre el aplicador de la prueba y el sujeto al que se le aplica; en el segundo, la relación se establece a través del hilo telefónico. La primera forma de aplicación suele utilizarse, por ejemplo, con niños pequeños, con personas que no entienden bien el idioma o con analfabetos. La segunda en los estudios de encuestas.

3. *Papel y lápiz*

   Tanto la presentación de la prueba como el registro de las respuestas de los sujetos se hacen en forma impresa.

4. *Mediante ordenador*

   Los ítems se van presentando en la pantalla del ordenador y los sujetos van respondiendo a cada uno de ellos a través del teclado. Actualmente este tipo de presentación está cobrando protagonismo gracias a los avances en el campo de la informática. Las ventajas de esta forma de aplicación hacen referencia tanto al menor coste de tiempo como a la mayor estandarización de las condiciones de administración y a las ventajas que ofrece el ordenador a la hora de registrar las respuestas, puntuarlas e interpretarlas (Olea y Hontangas, 1999).

5. *A través de correo*

   Esta forma de administración implica la desaparición de la figura del aplicador. La prueba se envía por correo (postal, electrónico), el sujeto que la recibe responde y la devuelve también mediante el mismo procedimiento. En general se suele enviar, junto a la prueba, una carta de saludo en la que se explica el objetivo del estudio y se solicita su colaboración, y un sobre convenientemente timbrado en el que está impresa la dirección a donde debe remitirse la prueba una vez cumplimentada. Se trata de una forma de administración bastante común en estudios de opinión y en aquellos que requieran la consulta de documentación para su cumplimentación. Tiene una ventaja, y es que de una manera muy sencilla se puede hacer un muestreo y enviar un gran número de pruebas para que sean contestadas. Sin embargo tiene algunos inconvenientes, uno de ellos, la alta tasa de personas que no responden (aproximadamente el 50%) y otro, la falta de seguridad de que la persona que responda sea aquella a la que se envió (Navas, 2002).

10. CORRECCIÓN DE LA PRUEBA PILOTO Y ASIGNACIÓN DE PUNTUACIONES A LOS SUJETOS

Una vez que se ha aplicado la prueba piloto, la primera tarea que ha de afrontar el investigador (el profesor, el educador....) es la de valorar las respuestas dadas por los sujetos a cada uno de los ítems para asignarles una puntuación. Esta tarea que puede parecer sencilla no lo es. Es necesario arbitrar la forma de que la puntuación asignada a cada sujeto refleje su nivel en la característica que se está midiendo y no otra cosa.

EJEMPLO:

La calificación obtenida en los exámenes por los alumnos de Psicometría debe reflejar únicamente los conocimientos que tienen de la asignatura y no debe depender, por ejemplo, del profesor que los haya corregido; si esto no fuera así, estaría influyendo en la calificación obtenida no sólo el nivel de los alumnos en la variable medida, sino los criterios seguidos por los profesores a la hora de corregir los exámenes. Esto haría imposible la comparación del nivel de los alumnos.

Es cierto que siempre que se emite un juicio acerca de algo es inevitable un cierto grado de subjetividad, pero también es cierto que es necesario tratar de eliminarla, o al menos controlarla, y para ello se dispone de distintos procedimientos. La elección de uno u otro dependerá del formato de los ítems que componen la prueba.

10.1. En los tests formados por ítems de elección

Este tipo de tests, también llamados de respuesta cerrada, tiene la ventaja de que el examinador no debe realizar ninguna valoración de las respuestas emitidas por los sujetos a cada uno de los ítems, eliminándose, por lo tanto, la posibilidad de introducir subjetividad en la puntuación que

se les asigne. Como se recordará, son los ítems utilizados en la mayoría de las pruebas de tipo cognitivo y en una gran parte de las elaboradas para la medida de variables no cognitivas.

10.1.1. En las pruebas cognitivas

En este tipo de pruebas, en las que hay respuestas correctas e incorrectas, para cada elemento se conoce de antemano cual es la respuesta correcta; por lo tanto, el proceso de corrección del test se reduce a comprobar si las respuestas emitidas por cada sujeto coinciden o no con las de una plantilla de corrección, asignando un uno por cada respuesta que coincida con la de la plantilla.

Una vez corregida la prueba, es necesario combinar las puntuaciones asignadas a cada elemento para obtener la puntuación de cada sujeto en el test total. La forma más habitual de proceder es sumar sencillamente el número de respuestas correctas.

$$\text{Puntuación} = \sum_{i=1}^{n} X_i \qquad [2.1]$$

Ahora bien, cuando se analizaron las ventajas derivadas del uso de este tipo de ítems, se vio también que tenían un inconveniente grave: la posibilidad de que un sujeto que desconociera por completo aquello que se le preguntaba eligiera por azar la respuesta correcta. Cuando un sujeto responde de esta manera, su puntuación final en el test será una estimación inflada de su verdadero nivel en el rasgo que se está midiendo. Por otra parte, si los sujetos no siguen el mismo patrón a la hora de responder, es difícil hacer comparaciones acerca de sus puntuaciones.

EJEMPLO:

Supongamos que en el examen de Psicometría hay dos alumnos que conocen 10 de las 20 preguntas que tiene el examen. Uno de ellos decide no arriesgarse y responde solamente a las 10 preguntas cuya respuesta conoce dejando las otras 10 en blanco. El otro alumno, más arriesgado, decide responder a todas las preguntas. Si las preguntas tienen dos alternativas de respuesta (verdadero-falso), ya comentamos que hay una probabilidad del 50% de que una persona que desconozca la respuesta correcta acierte por azar. En este caso, vamos a suponer que ha contestado correctamente a las 10 preguntas que conocía y que de las otras 10, al responder al azar, ha acertado el 50%, es decir, ha acertado 5 y ha fallado las otras 5. Este sujeto, que conocía el mismo número de preguntas que su compañero y, por lo tanto, debería haber obtenido la misma puntuación, al utilizar otro patrón de respuestas ha obtenido una mayor puntuación. El primero ha obtenido 10 puntos y el segundo 15.

Dado lo injusto del tema es necesario, o bien incidir en las instrucciones para que los alumnos no dejen ninguna respuesta en blanco, o bien utilizar algún procedimiento que permita controlar el efecto del azar sobre la puntuación final de los sujetos. Como no está claro el papel unificador de las instrucciones en la tendencia de los sujetos a responder al azar (Wood, 1987; Navas, 2002), es preferible utilizar una fórmula de corrección para llevar a cabo el control.

La aplicación de esta fórmula de corrección puede hacerse de dos maneras, o bien penalizando los errores cometidos, o bien bonificando las omisiones o ítems no respondidos.

1. Cuando se penalizan los errores es porque se asume que el sujeto no conoce la respuesta correcta y que todas las alternativas del ítem le resultan igualmente atractivas. Entonces las respuestas incorrectas son respuestas dadas al azar. Donde:

$$X_C = A - A_a = A - \frac{E}{K-1} \qquad [2.2]$$

X_c = puntuación corregida.

A = número de aciertos.

A_a = aciertos obtenidos al responder al azar.

E = número de errores.

K = número de alternativas de los ítems.

¿Cómo se obtiene esta fórmula de corrección?

Supongamos que el número de aciertos de un sujeto en el test viene dado por la puntuación A, pues bien, en esa puntuación están incluidos los aciertos que tuvo el sujeto porque conocía la respuesta y los que tuvo al responder al azar (A_a).

El valor de A_a no se puede calcular directamente, hay que inferirlo teniendo en cuenta el número de alternativas de respuesta. Si, como hemos apuntado antes, cuando un sujeto no conoce la respuesta correcta todas las alternativas son para él igualmente atractivas, la probabilidad de que elija por azar la respuesta correcta, es decir la probabilidad de acierto por azar, es $P(A_a) = \frac{1}{K}$, siendo K el número de alternativas. La probabilidad de que elija cualquiera de las otras opciones, es decir, la probabilidad de que cometa un error es: $P(E) = 1 - 1/K$ puesto que la suma de ambas probabilidades tiene que ser la unidad.

Si llamamos R_a al número de respuestas aleatorias que emite el sujeto en el total del test, se puede establecer que el número de errores será igual a:

$$E = R_a\left(1-\frac{1}{K}\right) = R_a\left(\frac{K-1}{K}\right)$$

es decir, será igual al número de respuestas aleatorias por la probabilidad de error.

El número de aciertos por azar será igual a:

$$A_a = R_a\frac{1}{K}$$

es decir, al número de respuestas aleatorias por la probabilidad de acertar por azar.

Despejando R_a en la fórmula de los errores tendremos:

$$R_a = E\frac{K}{K-1}$$

y sustituyendo en la fórmula de los aciertos por azar tendremos:

$$A_a = E\frac{K}{K-1}\left(\frac{1}{K}\right) = \frac{E}{K-1}$$

Si al número de aciertos totales le restamos el número de aciertos por azar, queda la fórmula de corrección tal y como la expusimos (ver 2.2).

Si aplicamos la fórmula de corrección a las puntuaciones obtenidas anteriormente por los dos alumnos en el examen de Psicometría vemos cómo al corregir el efecto del azar ambos obtienen la misma puntuación: $X = 15 - 5 = 10$

2. Cuando se bonifican las omisiones se parte del supuesto de que el sujeto sólo ha respondido a las preguntas que conocía, no ha respondido al azar a ninguna pregunta y por lo tanto no hay errores. En este caso, a la puntuación obtenida en el test se le añade una bonificación que correspondería a los aciertos que hubiere tenido si en lugar de dejar los ítems en blanco hubiera respondido al azar. La fórmula de corrección sería:

$$\boxed{X_C = A + A_a = A + \frac{O}{K}} \qquad [2.3]$$

Dado que no hay errores, el número de respuestas al azar coincidirá con el número de omisiones ($R_a = O$), y el número de aciertos al azar será el producto del número de omisiones por la probabilidad de acertar por azar ($A_a = O \cdot 1/K = O/K$).

Aplicando la fórmula a las puntuaciones obtenidas por los dos alumnos del ejemplo tendremos:

$$X_C = 10 + \frac{10}{2} = 15$$

Vemos que si se bonifican las omisiones al sujeto que no respondió al azar, ambos sujetos habrían obtenido también la misma puntuación.

Aunque también este procedimiento permitiría hacer comparaciones entre las puntuaciones de los alumnos, ambas puntuaciones estarían sobrevaloradas. No corresponderían al verdadero nivel de los sujetos, por lo tanto es más adecuado utilizar el primer procedimiento.

Cuando un mismo test está formado por ítems con distinto número de alternativas, para conocer cual es la puntuación de cada sujeto será necesario ir aplicando la corrección del azar por partes; se agruparán los ítems en función del número de alternativas y se calculará la puntuación del sujeto en cada uno de los grupos. La puntuación final será la suma de las puntuaciones parciales obtenidas.

EJEMPLO:

Supongamos un test formado por 100 ítems de los cuales hay 25 de dos alternativas de respuesta (verdadero-falso), 25 de 3 alternativas y 50 de 4 alternativas. ¿Cuál será la puntuación corregida de un sujeto que contestando a todos los ítems acertó 14 de verdadero-falso, 21 de los de tres alternativas y 29 de los ítems de 4 alternativas?

Si no se corrigiera el azar el sujeto obtendría una puntuación de 64 puntos sobre 100. En una escala de 10 puntos sería un 6,4.

Corrigiendo el efecto del azar tendríamos:

$$X_1 = 14 - \frac{11}{2-1} = 3$$

$$X_2 = 21 - \frac{4}{3-1} = 19$$

$$X_3 = 29 - \frac{21}{4-1} = 22$$

$$X_{total} = 3 + 19 + 22 = 44$$

El sujeto obtendría 44 puntos sobre 100, si se utilizara una escala de 10 puntos en lugar de una de 100, ese sujeto habría obtenido un 4,4.

10.1.2. En pruebas no cognitivas

En estas pruebas, en las que no hay respuestas correctas o incorrectas, los ítems llevan asignado un valor numérico distinto a cada alternativa de respuesta o categoría, lo que implica un escalamiento previo de los ítems (estímulos) en función del grado de atributo (o variable de interés) que manifiesten, o algún tipo de codificación previa. Entonces la forma de corregir el test y asignar puntuaciones a los sujetos suele ser sumando los valores numéricos asignados a las alternativas o categorías de respuesta elegidas por el sujeto. Esto nos hace pensar en la necesidad de que la asignación numérica a cada categoría de respuesta y a cada ítem esté bien hecha.

¿Cuál es el problema? Pues que cuando se utiliza un formato de escalas de categorías o clasificación, por ejemplo, hay que tener muy claro cuál es la dirección del continuo de la variable que se está midiendo. Si se trata de una variable de actitud, hay que conocer cuál es el extremo del continuo que marca una actitud favorable y cuál es el que marca una actitud desfavorable. Si es un test para medir depresión, se deberá saber cuál es el extremo que indica falta de depresión y cuál el que hace referencia a un grado máximo de depresión. Una vez aclarado este punto, es necesario decidir a qué extremo del continuo se le va asignar el valor numérico más alto y, finalmente, tener cuidado de que en todos los ítems del test se siga la misma regla de asignación.

Hay diferentes procedimientos para asignar los valores numéricos a los ítems o a las distintas categorías de respuesta de cada uno de ellos; en el tema 3 se abordará el estudio de los más utilizados y los principios en los que se basan. El ejemplo que aparece a continuación corresponde a una escala tipo Likert.

EJEMPLO:

Debería prohibirse beber alcohol en los sitios públicos:

1. Totalmente en desacuerdo
2. En desacuerdo
3. Me es indiferente
4. De acuerdo
5. Totalmente de acuerdo

El alcohol es bueno tomado con moderación

5. Totalmente en desacuerdo
4. En desacuerdo
3. Me es indiferente
2. De acuerdo
1. Totalmente de acuerdo

De los dos ítems, el primero muestra una actitud marcadamente contraria al consumo del alcohol. Si se asigna el valor numérico tal y como se ha hecho, el 1 correspondería al extremo que representa una actitud más favorable hacia el alcohol mientras que el 5 correspondería al extremo que representa una actitud más desfavorable. Si esto se hace así con un ítem hay que hacerlo con todos los demás. Por eso en el segundo ítem cuyo enunciado denota una actitud más positiva hacia el consumo de alcohol, la asignación de los valores numéricos se ha invertido de manera que el valor numérico máximo represente una actitud desfavorable hacia el consumo de alcohol.

De esta manera, cuando se corrija la prueba completa, los sujetos que muestren una actitud más favorable hacia el consumo de alcohol obtendrán valores más bajos que aquellos que muestren una actitud desfavorable.

10.2. En los tests formados por ítems de construcción

Dentro de este tipo de pruebas, las formadas por *ítems de respuesta corta* no presentan demasiado problema, cuando se trata de que el sujeto construya la respuesta correcta con una palabra o frase corta es fácil la asignación de la puntuación a los sujetos. El problema se va complicando a medida que las respuestas son más abiertas y extensas puesto que es más difícil controlar la subjetividad en la corrección. La persona que las corrija ha de emitir juicios valorativos acerca de la adecuación de las respuestas.

Como ya apuntamos anteriormente, este tipo de pruebas presentan la ventaja, sobre las pruebas objetivas, de que permiten expresarse abiertamente a los sujetos, y de esta manera se pueden valorar no sólo los conocimientos del tema, sino otros aspectos importantes en algunas situaciones como: la forma de redactar, la creatividad, la forma de estructurar el tema, la capacidad para resumir y esquematizar, y un largo etc. Pero tienen el grave inconveniente de que la corrección de las mismas, además de muy laboriosa, es bastante subjetiva. No obstante hay formas de controlar y reducir esa subjetividad: *Método de la puntuación analítica y Método de la puntuación holística.*

10.2.1. Método de la puntuación analítica

Este método requiere, en primer lugar, definir de forma inequívoca y aislar las dimensiones que se consideran importantes para la realización de la tarea a evaluar y, una vez establecidas las dimensiones que hay que considerar en la corrección de la prueba, es necesario establecer la forma de evaluarlas, definiendo claramente lo que se considera una respuesta adecuada o correcta en cada dimensión y estableciendo el número de respuestas correctas que se necesitan, en cada una de ellas, para poder decir que la tarea ha sido correctamente realizada. Si realmente se llegan a definir claramente estos criterios, las pruebas pueden ser corregidas por personas que no sean expertas en la materia a evaluar ya que no habrá dificultad en decidir si una respuesta es correcta o no. Con este procedimiento la puntuación final de los sujetos suele venir expresada mediante dos únicos valores: correcta/incorrecta, apto/no apto, aprobado/suspenso, etc. pero se obtiene información de cada una de las dimensiones.

10.2.2. Método de la puntuación holística

En este procedimiento se evalúa de una manera global u holística la forma en que los sujetos han realizado la prueba, y la puntuación asignada, que podrá tomar distintos valores dentro de los límites establecidos de antemano, expresará la calidad global de su respuesta. A diferencia del anterior, este procedimiento requiere que la corrección de las pruebas sea hecha por expertos en la materia a evaluar previamente entrenados para tratar, en lo posible, de alcanzar un acuerdo entre ellos y eliminar la subjetividad en la corrección.

¿Qué método es más adecuado? Todo dependerá del objetivo para el que se construyó el test. En algunos casos se requerirá una información más pormenorizada acerca de la ejecución de los sujetos en la prueba, por ejemplo si lo que se quiere es detectar déficits o hacer algún diagnóstico; pero otras veces bastará con tener una información global, por ejemplo en los exámenes destinados a la obtención de una titulación académica, o de competencia profesional.

11. EJERCICIOS DE AUTOEVALUACIÓN

A continuación se presentan una serie de ítems. Unos estarán correctamente redactados y otros presentarán distintos errores. Después de leerlos atentamente responda indicando el tipo de formato que tiene cada uno de ellos, si están correctamente redactados y, en caso contrario, cómo deberían haberlo estado. Decir también si la variable que miden pertenece al ámbito cognitivo o, por el contrario, al ámbito oréctico o afectivo.

1. _____ es a vino como trigo es a _____

A. uva – avena

B. agua – pan

C. uva – harina

D. beber – comer

E. agua – avena

2. Determinar el número que falta (indicado con puntos suspensivos) para que resulte correcta la siguiente igualdad:

(12 ×.....) – 6 = 3

A. 0,075

B. 0,0075

C. 0,00075

D. 0,75

3. Actualmente no encuentro muy difícil no perder la esperanza de no llegar a ser algo _____ Verdadero Falso

4. Tengo opiniones políticas sólidas

1. Completamente de acuerdo
2. De acuerdo
3. No se
4. En desacuerdo
5. Completamente en desacuerdo

5. A continuación se presenta un término de la lengua castellana y cinco definiciones de las que sólo una es correcta. Leer las cinco opciones de respuesta y elegir la opción correcta.

LIPOTIMIA

A) Máquina de componer que contiene todas las letras de una línea

B) Desmayo con pérdida de sentido

C) Son ciertas la A y la B

D) Es cierta la B

E) Acumulación de grasa en la piel

6. El cuadro de las Meninas fue pintado por _____

7. Me despierto nervioso por las mañanas

1. Siempre
2. Casi siempre
3. A veces
4. Casi nunca
5. Nunca

8. A continuación se presentan dos columnas. En la columna de la izquierda, la de las premisas, se incluyen los títulos de 5 cuadros famosos. En la columna de las respuestas se recogen los nombres de 5 pintores. Unir mediante flechas el pintor con su cuadro correspondiente:

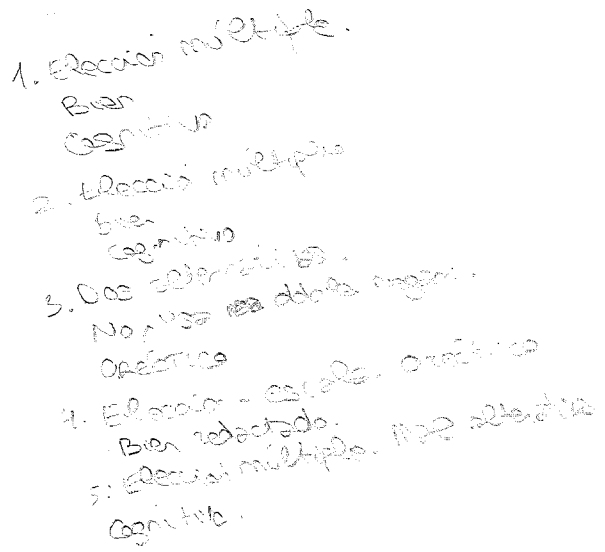
A. El nacimiento de Venus	1. Velázquez
B. El entierro del Conde Orgaz	2. Botticelli
C. La Venus del espejo	3. El Greco
D. La adoración de los pastores	4. Goya
E. La Maja desnuda	5. Rembrandt

9. Ejercicios conceptuales

Después de leer detenidamente el enunciado que se presenta, el lector deberá responder si es verdadero o falso y justificar su respuesta.

1. La etapa de definición de la finalidad del test es la etapa en la que se decide el tipo de formato de los ítems.
2. Los ítems de elección múltiple son ítems de respuesta abierta.
3. En los tests de velocidad el tiempo de ejecución está limitado.
4. Los tests de potencia son típicos de pruebas de aptitudes.
5. Los ítems de ensayo son ítems de respuesta abierta.
6. A medida que aumentan las opciones de respuesta en los tests de elección múltiple, disminuye la probabilidad de acierto por azar.
7. En los listados, las distintas opciones de respuesta están ordenadas de forma graduada.

8. Los ítems de elección múltiple se utilizan sobre todo para medir variables de tipo cognitivo.
9. Los tests de potencia, aplicados a la población general, deben estar formados por ítems muy fáciles.
10. Si se quiere que un test sirva para detectar a los niños que tienen un menor nivel de conocimientos los ítems deberán ser fáciles o muy fáciles.



12. SOLUCIONES A LOS EJERCICIOS DE AUTOEVALUACIÓN

1. Se trata de un ítem de elección en el que la tarea del sujeto será completar los espacios en blanco con alguna de las opciones de respuesta que se le ofrecen. Es de tipo cloze (o de completar) y está correctamente presentado. La variable que mide es de tipo cognitivo y la alternativa correcta es la C.

2. Se trata también de un ítem de elección, de las mismas características que el anterior, en el que la tarea del sujeto consiste en elegir de entre las alternativas aquella que complete la ecuación y verifique la igualdad. La solución correcta es la D.

3. Se trata de un ítem muy mal redactado ya que tiene muchas negaciones. Es imposible interpretar lo que quiere decir y, por lo tanto, responder. Se trata de un ítem de elección binaria en el que hay dos respuestas de las cuales se supone que una es verdadera. La variable que se intenta medir no pertenece al ámbito cognitivo, intenta medir opiniones.

4. Es un ítem cuyo formato corresponde a una escala de clasificación en la que las respuestas están ordenadas formando una escala graduada a lo largo del continuo de la variable que se quiere medir, en este caso las opiniones políticas. Por lo tanto, no es una variable de tipo cognitivo.

5. Se trata de un ítem de elección múltiple que mide conocimientos, pero está muy mal formulado ya que, a pesar de que dice que sólo hay una respuesta correcta, hay dos, la B y la D. Por otra parte, como ya se comentó a lo largo del tema, hay que procurar que las opciones de respuesta sean lo más independientes posible entre sí y evitar que las alternativas sean del tipo: A y B. Este tipo de alternativas provocan ruido en los sujetos.

6. Es un ítem de construcción, de respuesta corta, que mide conocimientos. Está bien formulado y la tarea del sujeto consistirá en rellenar el espacio en blanco con el nombre del pintor. En este caso Velázquez.

7. Se trata de un ítem de elección, de respuesta cerrada, que mide una variable no cognitiva y está bien redactado. La tarea del sujeto será elegir la categoría que mejor represente su estado.

8. Se trata de un ítem de respuesta cerrada, de emparejamiento, que mide conocimientos. Está bien planteado y la tarea del sujeto será elegir de la columna de la derecha el pintor que corresponda a cada una de las obras situadas en la columna de la izquierda y unir ambos elementos mediante flechas. En este caso habría que unir: (A, 2), (B, 3), (C,1), (D,5) y (E,4).

9. Soluciones a los ejercicios conceptuales:

1. La afirmación es falsa

   Es en la etapa de especificación de las características del test donde se decide acerca del formato que van a tener los ítems. La definición de la finalidad del test es una etapa previa a la construcción propiamente dicha, en ella se ha de decidir acerca de la variable a medir, de la población a la que va dirigido el test y del uso que se va a dar al test.

2. La afirmación es falsa.

   Los ítems de elección múltiple son de respuesta cerrada. Junto al enunciado del ítem se ofrecen una serie de alternativas de respuesta entre las que se deberá elegir aquella que se considere la correcta o la más correcta.

3. La afirmación es verdadera.

   Los ítems que forman los tests de velocidad son lo suficientemente fáciles como para que los contestaran correctamente todos los sujetos si dispusieran de tiempo suficiente. La forma de discriminar entre los sujetos es, precisamente, limitar el tiempo para la ejecución de la prueba.

4. La afirmación es verdadera.

   Los tests de potencia están formados por ítems de distinta dificultad y tratan de medir el nivel de conocimientos o aptitudes de los sujetos.

5. La afirmación es verdadera.

   En los ítems de ensayo el sujeto debe elaborar su propia respuesta.

6. La afirmación es correcta.

   Si todas las alternativas de respuesta son igualmente atractivas para un sujeto que desconoce la alternativa correcta, la probabilidad de acierto por azar es igual a 1/K, siendo K el número de alternativas. De ahí se desprende que a medida que aumenta el número de alternativas disminuye la probabilidad de acertar la correcta por azar.

7. La afirmación es falsa.

   Los listados, se diferencian de las escalas de clasificación en que las opciones de respuesta no forman una escala ordinal graduada y se diferencian de los ítems de elección múltiple en que no hay respuestas correctas o incorrectas. Junto al enunciado del ítem se ofrece una lista de posibles respuestas entre las que deberá elegir el sujeto aquella o aquellas con las que esté de acuerdo. Las opciones de respuesta son independientes entre sí.

8. La afirmación es correcta.

   Este tipo de formato es el más adecuado cuando se quiere obtener una medida objetiva de variables cognitivas.

9. La afirmación es incorrecta.

   Dado que se trata de discriminar entre los sujetos y estos disponen de tiempo suficiente para responder, los ítems deben cubrir todo el continuo de dificultad, desde ítems muy fá-

ciles que sólo los menos capacitados respondan de forma incorrecta, hasta ítems muy difíciles que sólo los más capacitados los puedan acertar.

10. La afirmación es correcta.

    En la pregunta anterior ya se contestó, en parte, a esta pregunta. Ahora bien, si lo que nos interesa es discriminar sólo en este sector de la población, no es necesario incluir elementos con diferentes niveles de dificultad, todos los ítems pueden ser fáciles o muy fáciles; entonces los responderán correctamente todos los sujetos de la muestra a excepción de los menos capacitados.

13. BIBLIOGRAFÍA COMPLEMENTARIA

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

En el capítulo 2, dedicado a la construcción de un test, ofrece una panorámica muy general acerca de la forma de llevar a cabo el proceso.

Navas, M.J. (2002). *Métodos, diseños y técnicas de investigación psicológica*. Madrid: UNED.

En el capítulo 3 ofrece una revisión muy completa y clara ya que utiliza un lenguaje muy sencillo y adaptado al nivel de conocimientos de nuestros alumnos.

Thorndike, R.L. (1989). *Psicometría aplicada*. Méjico: Limusa

Los capítulos 2, 3 y 4 están dedicados al tema que nos ocupa.

TEMA 3

TÉCNICAS PARA LA CONSTRUCCIÓN DE ESCALAS DE ACTITUDES

María Isabel Barbero García

SUMARIO

1. ORIENTACIONES DIDÁCTICAS

En el tema anterior, se hizo una introducción general a los principios básicos y a las normas que han de guiar la construcción de instrumentos de medición psicológica. Creemos que el tema incluye una información bastante completa y clara, no sólo de cara a la elaboración de tests para la medición de variables cognitivas: aptitudes, rendimiento, conocimientos, etc.; sino para la construcción de escalas, cuestionarios o inventarios que permitan la medición de variables no cognitivas: personalidad, actitudes, intereses, valores, opiniones, etc. Sin embargo, dada la importancia que para el psicólogo tienen este tipo de pruebas y lo habitual de su uso, hemos creído necesario dedicar un tema de este libro a la descripción de las principales técnicas desarrolladas para la elaboración de escalas que permitan la medición de estas variables. Por otra parte, sería imperdonable que nuestros alumnos desconocieran las aportaciones de figuras tan importantes como Thurstone, Likert, Osgood y Guttman.

Aunque el tema se centra en la construcción de escalas para la medición de las actitudes, las técnicas expuestas se pueden adaptar para su utilización en la medición de intereses y valores, entre otras muchas variables.

Nunnally (1978) hace la siguiente distinción entre intereses, valores y actitudes:

Los *Intereses* son preferencias por actividades particulares. Por ejemplo, dos ítems relacionados con intereses podrían ser:

— Prefiero reparar un coche a leer un libro.

— Prefiero trabajar al aire libre que en un despacho.

Se han llevado a cabo numerosas investigaciones sobre intereses, sobre todo intereses vocacionales, de cara a la orientación vocacional.

Los *Valores* hacen referencia a preferencias sobre objetivos de vida y formas de vida más que sobre actividades concretas. Por ejemplo dos ítems que serían adecuados en una escala de valores serían:

$$C.A. = Q_3 - Q_1 = 10,25 - 8,13 = 2,12$$

Si el coeficiente de ambigüedad es mayor que 2, el elemento será considerado ambiguo y deberá eliminarse de la escala definitiva. En elementos neutrales; es decir, en aquellos cuyo valor escalar esté comprendido en el intervalo 5,5 - 6,5 si la escala tiene 11 categorías, o en el punto central de la escala sea cual sea el número de categorías, el coeficiente de ambigüedad puede llegar a 3. En nuestro caso, tanto el ítem 5 como el 6 tienen un coeficiente de ambigüedad algo superior a 2 por lo que, en teoría, habría que eliminarlos de la escala, pero al estar muy próximos al valor 2 se podrían mantener en la escala.

Por otra parte, la escala completa debe incluir ítems que cubran todo el continuo de actitud, desde el extremo más desfavorable al más favorable.

Sea cual sea el procedimiento utilizado (comparaciones binarias, intervalos aparentemente iguales...), una vez asignados los valores escalares a los ítems, la escala de actitud ya está lista para ser utilizada. Se cuenta ya con un instrumento de medida, una escala, que podrá ser aplicada a una muestra piloto de sujetos para su evaluación y construcción de la escala definitiva que permitirá medir la actitud de los sujetos ante la variable objeto de estudio.

Los ítems pueden ordenarse, para su presentación a los sujetos, bien de forma aleatoria o en función de sus valores escalares. La respuesta de los sujetos a cada uno de ellos es una respuesta dicotómica. Se indica a los sujetos que lean detenidamente el enunciado de cada ítem y que, en función de *su posición personal, sus propios sentimientos, opiniones, o su propia actitud* ante cada uno de ellos, respondan si están de acuerdo con su enunciado o no. Es decir, ahora los sujetos no deben emitir juicios de hecho como tenían que hacer los jueces, sino juicios de valor.

La puntuación en la escala para cada sujeto se obtiene calculando la media de los valores escalares de los ítems con los cuales el sujeto estuvo de acuerdo.

Supongamos que un sujeto ha mostrado su acuerdo con cuatro ítems de la escala que estábamos construyendo para medir la actitud hacia el matrimonio, y que los valores escalares de esos ítems son: 8,5; 9,3; 10 y 8,7 respectivamente; la puntuación de ese sujeto en la escala será la media de esos valores; es decir: (8,5 + 9,3 + 10 + 8,7)/4 = 9,12. Este valor indica que la actitud del sujeto es bastante favorable a la institución matrimonial.

La principal ventaja de las escalas de actitudes de Thurstone sobre otro tipo de escalas es que permiten la interpretación directa de la actitud de un sujeto sin necesidad de hacer referencia al grupo, o la actitud media de un grupo de sujetos sin recurrir a normas generales. Sin embargo en la mayoría de los estudios en Psicología y Sociología esto no es realmente una ventaja. En la mayoría de los estudios el investigador está interesado en correlacionar las diferencias individuales en actitud con otro tipo de diferencias individuales, o está interesado en el estudio de las diferencias de actitud entre distintos grupos. En este caso no tiene necesidad de una interpretación directa de la actitud de una persona concreta. Aún en casos en que este tipo de interpretaciones sea importante, las escalas de Thurstone deberán utilizarse con ciertas precauciones debido a que a veces puede suceder que los valores asignados a los ítems dependan en gran medida de los jueces que se utilicen.

Hoy día se considera que los modelos sumativos, como el que veremos a continuación desarrollado por Likert, son más útiles para la medida de las actitudes.

3. LA TÉCNICA DE LIKERT

Esta técnica surge para tratar de dar una solución razonable al problema que se planteaba en relación con los aspectos cuantitativos del estudio de las actitudes sociales, y su origen hay que situarlo en una investigación iniciada por Gardner Murphy en 1929.

Likert consideraba que el método de Thurstone era muy laborioso ya que incluía, entre otras pruebas, la *prueba de jueces*; entonces, se plantea la posibilidad de elaborar un tipo de escalas más sencillas pero igualmente fiables, en las que no hubiera necesidad de utilizar tantas comprobaciones estadísticas.

La técnica propuesta es el *modelo sumativo* más utilizado para la medida de las diferencias individuales respecto a los rasgos psicológicos. En ella, sólo se asume que los ítems están monotónicamente relacionados con el rasgo subyacente que se quiere medir; es decir, que a medida que aumenta o disminuye la cantidad de rasgo manifestado por los sujetos, aumenta o disminuye su puntuación en el ítem y que la suma de las puntuaciones a los ítems está relacionada linealmente con el rasgo. La puntuación total se obtiene sumando las puntuaciones de los sujetos a cada uno de los ítems (teniendo en cuenta el valor asignado a cada ítem) y tienen la ventaja de que son fáciles de construir, son muy fiables, pueden ser adaptadas para medir cualquier tipo de actitud y han producido resultados significativos en distintos estudios.

La forma de construir escalas de Likert es un caso especial del método general de construcción de tests de potencia (no de velocidad).

Las escalas resultantes están incluidas entre las estudiadas en el tema anterior como escalas de clasificación o de categorías, al igual que las escalas de Thurstone elaboradas mediante la ley del Juicio Categórico.

3.1. Fundamentos de la técnica

Likert parte del supuesto de que las actitudes pueden medirse a través de las manifestaciones verbales de los sujetos, y la técnica que propone para la medida de las actitudes se basa en los siguientes principios y postulados recogidos por López Pérez, 1985, pág. 251:

1. Es posible estudiar dimensiones de actitud a partir de un conjunto de enunciados que operen como reactivos para los sujetos.
2. Los individuos pueden situarse en la variable de actitud desde el punto más favorable al más desfavorable. La variación de las respuestas será debida a diferencias individuales de los sujetos.
3. La valoración de los sujetos en la variable de actitud no supone una distribución uniforme sobre el continuo de actitud, sino su posición favorable o desfavorable sobre el objeto estudiado.

La técnica de Likert surge cuando en 1929 junto con Gardner Murphy se propone presentar un amplio conjunto de problemas relacionados con determinadas áreas de actitudes: relaciones internacionales, problemas raciales, conflicto económico, religión, etc. partiendo del supuesto de que las actitudes sociales se agrupaban en pautas. Bajo ese supuesto, si se podía contar con una serie de ítems que hicieran referencia a un mismo problema social general, y se conociera la actitud de un sujeto frente a algún aspecto de dicho problema, se podría predecir la actitud o actitudes que manifestaría dicho sujeto respecto a otros aspectos del mismo problema.

Desde el punto de vista de la medición, la técnica de Likert asume un nivel de medida ordinal. Los sujetos son ordenados en la escala en función de su posición favorable/desfavorable respecto a la actitud medida.

Se trata, además, de una *escala sumativa* ya que la puntuación obtenida por los sujetos en la escala es función de las puntuaciones obtenidas en cada uno de los ítems o elementos que la componen. Esto implica dos supuestos adicionales:

1) Que la suma de las curvas características de los ítems sea una función monotónica y aproximadamente lineal respecto a la actitud medida.
2) Que todos los elementos que componen la escala estén midiendo una única dimensión. Se trataría por lo tanto de una escala unidimensional.

¿Qué significa que las curvas características de los ítems sean monotónicas respecto a la actitud medida?

Vamos a explicarlo con un ejemplo:

Cuando se utilizan tests objetivos para la medición de las aptitudes, partimos del supuesto de que cuanto mayor sea la capacidad o aptitud de una persona, mayor será la probabilidad de que responda correctamente a una determinada pregunta o elemento del test. Por lo tanto, aquellos elementos que sean contestados incorrectamente por sujetos que tienen una aptitud alta y correctamente por los que tienen aptitud baja deberían ser eliminados en el proceso de selección, ya que sus curvas características no son función monotónica creciente; pues lo mismo ocurre cuando se trata de medir actitudes mediante una escala de Likert, cuanto más favorable sea la actitud de un sujeto hacia aquello que se está midiendo, mayor será la probabilidad de que elija en cada ítem

la categoría que indique esa postura. No es normal que sujetos que muestran una actitud muy favorable hacia aquello que se está midiendo, elijan ítems que representen actitud desfavorable; si esto ocurriera deberían ser eliminados esos ítems de la escala definitiva o revisar si la asignación de puntuaciones a las distintas categorías está bien hecha.

Nota: Una exposición más detallada se encontrará en el tema 8 al hacer el análisis de la calidad métrica de los ítems.

La redacción y presentación de los ítems ha de permitir a los sujetos emitir *juicios de valor* y no *juicios de hecho*, es decir, ante cada uno de los ítems los sujetos deben expresar lo que según ellos *debería ser* no lo que *de hecho sea*.

Un ejemplo podría ser:

La familia debería permanecer más tiempo reunida

Ante este enunciado, los sujetos deberán responder eligiendo entre una serie de categorías aquella que mejor se adapte a su postura personal.

Como puede apreciarse, hay una clara diferencia en este tipo de escala respecto a la prueba de jueces de las escalas de Thurstone; en éstas se les pedía a los jueces que no emitieran *juicios de valor*, sino que emitieran *juicios de hecho*. Una vez construida la escala, cuando se aplica a los sujetos para evaluar su actitud, entonces éstos deberán emitir los *juicios de valor*.

Dado que, en las escalas de Likert, lo que se piden son juicios de valor de los sujetos, cada problema debe ser presentado de forma que cada sujeto pueda tomar partido entre alternativas opuestas.

La forma de responder a los ítems puede variar, aunque normalmente los sujetos han de responder en función de cinco categorías:

a) Completamente de acuerdo
b) De acuerdo
c) Indiferente
d) En desacuerdo
e) Completamente en desacuerdo

3.2. Asignación de valores numéricos a los ítems y puntuaciones a los sujetos

Aunque ya se ha explicado en el tema anterior vamos a recordarlo. Una vez que se han redactado los ítems, hay que analizar si su enunciado representa una actitud positiva o negativa respecto a la ac-

titud que se quiere medir y, después de evaluar este aspecto, hay que asignar un valor numérico a cada una de las opciones o categorías de respuesta. Esa asignación se deja al arbitrio del investigador, pero debe de ser hecha de forma que se mantenga la coherencia interna en el sentido de la actitud medida. Es decir, es necesario que siempre el valor más alto indique una actitud más positiva hacia aquello que se está midiendo. El número de opciones depende de lo que pretenda el investigador, de la naturaleza de la variable a estudiar y del tipo de elementos o ítems que se estén utilizando. Las escalas de Likert utilizan, normalmente, cinco opciones de respuesta, pero se puede utilizar otras.

Supongamos por ejemplo el elemento comentado anteriormente:

La familia debería permanecer más tiempo reunida

Si utilizamos cinco categorías de respuesta para la evaluación de este elemento, la asignación de valores numéricos a esas categorías podría ser:

Completamente en desacuerdo	1
En desacuerdo	2
Indiferente	3
De acuerdo	4
Completamente de acuerdo	5

Otra forma de puntuar las categorías sería:

Completamente en desacuerdo	-2
En desacuerdo	-1
Indiferente	0
De acuerdo	1
Completamente de acuerdo	2

La puntuación de los sujetos en la escala total, será la suma de los valores numéricos asignados a cada una de las categorías elegidas por los sujetos en el conjunto de los ítems.

4. EL DIFERENCIAL SEMÁNTICO DE OSGOOD

Se trata de una escala de clasificación elaborada por Osgood y sus colaboradores (1957), con el fin de medir el significado connotativo, también llamado significado afectivo o subjetivo, que determinados estímulos tienen para los sujetos. Osgood estaba interesado en las reacciones emocionales que las palabras o conceptos producen en las personas.

Durante mucho tiempo los filósofos y lingüistas han estado preocupados por el estudio del significado de las palabras, frases, etc.; sin embargo, los psicólogos, a pesar de su interés por desentrañar la naturaleza del lenguaje y de los procesos de comunicación, tardaron más en ponerse a trabajar, sistemáticamente, en la elaboración de teorías del significado y en la investigación empírica del fenómeno.

Noam Chomsky y los seguidores de su obra desarrollaron, dentro del campo de la lingüística, las teorías estructuralistas del significado, dando origen a un campo de investigación muy amplio que ha permitido a los psicólogos abordar el problema del lenguaje y la naturaleza del mismo sobre una base firme.

Hay varias formas de aproximación al problema del significado, entre las que podemos citar: las teorías estructurales y las teorías del proceso mental.

Osgood hizo una revisión sistemática de todas las teorías del significado y así pudo encontrar el marco teórico que le permitió desarrollar un instrumento para medirlo: El Diferencial Semántico.

Partió de la consideración de que la actitud que una persona muestre hacia un objeto dependerá del significado evaluativo que dicho objeto tiene para la persona. Por otra parte, como recoge Visauta (1989, pág. 220), el principio fundamental en el que se basa el diferencial semántico es que la gran diversidad de significados es reducible a unas determinadas variaciones en un número limitado de dimensiones.

El campo de aplicación del Diferencial Semántico es muy amplio ya que, debido a su naturaleza y a su adaptabilidad, se ha convertido en un instrumento de medida muy utilizado. Dentro de la Psicología podemos hablar de cuatro áreas en las que su uso es habitual: en la investigación clínica, en la medida de las actitudes, en investigaciones transculturales y en investigaciones sociales (Tomado de Díaz- Guerrero y Salas, 1975).

Aunque a partir del nombre: *El Diferencial Semántico* podría inferirse que se trata de una prueba formada por unos ítems concretos como puede ser, por ejemplo, el *Test de Matrices Progresivas* de Raven o cualquier otro test; en realidad se trata de una forma distinta de abordar el problema de la medida de las actitudes.

El formato de la escala consiste en la presentación a los sujetos de un *concepto* seguido de una serie de *escalas* cuyos extremos están marcados por adjetivos bipolares.

De todo lo anterior se desprende que hay dos elementos fundamentales en el Diferencial Semántico (D.S.): los conceptos y las escalas bipolares.

4.1. Los Conceptos

El término concepto tiene aquí un sentido amplio, ya que hace referencia al estímulo u objeto que ha de evaluar el sujeto.

Los estímulos pueden ser de lo más variado y aunque, en general, se refieren a conceptos verbales (Dios, madre, educación, acciones políticas, etc.), se pueden referir a conceptos no verbales (cuadros, esculturas, estímulos físicos, etc.) por eso, en primer lugar, hay que definir claramente el problema o área a investigar y, posteriormente, elegir los conceptos más adecuados para llevar a cabo la investigación.

Osgood utilizó en la mayoría de sus investigaciones sustantivos como estímulos, aunque también usó adjetivos como él mismo nos indica en el capítulo 3 de su libro *The Measurement of Meaning*.

Estos conceptos aparecerán encabezando el formulario, como se explicará posteriormente, seguidos del conjunto de escalas bipolares que se utilizarán para llevar a cabo dicha evaluación.

Dado que es prácticamente imposible cubrir, a base de conceptos, todo el área a investigar, es necesario hacer un muestreo de todo el universo de conceptos que la definen para extraer aquellos que sean más relevantes y representativos; no obstante, como afirma Osgood, a veces el investigador se guía por su «buen juicio» y tiende a elegir aquellos conceptos que:

a) Discriminan bien entre los sujetos ya que, de esta manera, se obtiene una mayor información.

b) Tengan un significado claro y único para el sujeto, de manera que cuando se le presenten sepa lo que está juzgando.

c) Sean familiares a todos los sujetos de la muestra para que la respuesta que ofrezcan sea real y no esté sesgada debido a la falta de familiaridad con el concepto a evaluar. A través de sus experiencias Osgood encontró que, cuando los sujetos no están familiarizados con el concepto que se está evaluando, se produce una regresión hacia el punto medio en la escala de evaluación.

4.2. Escalas bipolares

El significado de los conceptos (estímulos) se evalúa por medio de escalas semánticas bipolares.

Cada una de estas escalas bipolares representan una reacción de tipo afectivo hacia el objeto: Bueno-Malo, Sano-Enfermo, etc., y lo que se pretende es utilizarlas de manera que se pueda obtener una medida del significado afectivo que cada objeto (estímulo) tiene para los sujetos.

Estas escalas están ancladas en sus extremos por dos adjetivos antónimos, que describen un aspecto del continuo semántico: Fuerte-Débil, Grande-Pequeño, etc., a lo largo del cuál se situará el concepto evaluado. En general, el continuo se encuentra dividido en siete categorías, aunque se puede utilizar otro número, y la tarea del sujeto será evaluar el concepto y clasificarlo en función de la relación que haya entre éste y uno de los polos de la escala; para ello, pondrá una marca en el punto del continuo donde crea que debe situarse el concepto.

Supongamos, por ejemplo, que se ha pedido a un sujeto que evalúe el concepto MADRE, y una de las escalas bipolares que tiene para hacer la evaluación es: ACTIVA-PASIVA; pues bien, si a través de la evaluación subjetiva que haga dicho sujeto del concepto MADRE considera que la mejor representación de su significado es *muy activa*, habrá de colocar una marca en la categoría más próxima al adjetivo: *activa*; mientras que, si fuera *muy pasiva*, lo deberá hacer en la categoría más próxima al adjetivo: *pasiva*. Entre ambos extremos estarán los grados intermedios. Cuando el concepto tiene para el sujeto un significado neutro o indiferente, colocará su marca en la categoría central.

EJEMPLO:

POLITICA

Mala	1 2 3 4 5 6 7	Buena
Inútil	1 2 3 4 5 6 7	Útil
Deshonesta	1 2 3 4 5 6 7	Honesta
Injusta	1 2 3 4 5 6 7	Justa
Necia	1 2 3 4 5 6 7	Sabia

Los números asignados a cada una de las escalas son los que van a permitir obtener una escala sumativa que represente la evaluación que el sujeto ha hecho del concepto en cada una de ellas.

Todas las escalas bipolares que se han utilizado en el ejemplo hacen referencia a una *dimensión evaluativa* del concepto, pero hay otros pares de adjetivos que hacen referencia a otro tipo de dimensiones, por ejemplo *de potencia o de actividad*, como veremos más adelante.

Cuando un sujeto clasifica un concepto en la categoría media de la escala, diremos que considera que no hay asociación ni relación semántica entre el concepto y la escala bipolar utilizada. Este tipo de respuestas, como hemos comentado anteriormente, se pueden obtener si el concepto a evaluar no tiene un significado familiar para los sujetos.

La forma de presentación del Diferencial Semántico es muy variada; en general, se utiliza como una prueba de papel y lápiz y así la aplicación puede ser colectiva y se pueden evaluar varios conceptos a la vez. En este caso, aparecerán cada uno de los conceptos a evaluar seguidos de sus escalas bipolares correspondientes.

4.3. El espacio semántico: criterios de selección de las escalas

El número de escalas bipolares que se puede utilizar para evaluar un concepto determinado es prácticamente ilimitado de ahí que, a la hora de hacer una selección de las mismas, debamos tratar de obtener las más representativas. Ahora bien, ¿qué entendemos por las más representativas?, ¿representativas de qué? Estas serían dos de las posibles preguntas a las que habremos de dar respuesta.

Podemos considerar que el significado semántico de cualquier concepto está definido por una serie de dimensiones, subyacentes al mismo, que hemos de evaluar por medio de las escalas bipolares; de ahí que lo que tratamos de decir al hablar de seleccionar las escalas más representativas es, precisamente, subrayar la necesidad de utilizar las que mejor vayan a medir las dimensiones subyacentes al significado semántico del concepto a evaluar.

Cuando Osgood y sus colaboradores elaboraron el D.S. utilizaron 40 estímulos, y para conseguir las escalas que se iban a utilizar para evaluarlos, se pidió a una muestra de 200 estudiantes que emitieran, ante cada uno de los estímulos, una lista de aquellos adjetivos que consideraban que podían aplicárseles. Posteriormente, se analizaron las listas obtenidas y se seleccionaron aquellos adjetivos que habían aparecido con mayor frecuencia, buscando a continuación su opuesto para, de esa forma, obtener las escalas bipolares. La escala definitiva estuvo compuesta por 20 conceptos con 50 escalas bipolares cada uno de ellos.

Actualmente, contamos con numerosas técnicas estadísticas que nos van a permitir identificar y aislar las dimensiones subyacentes al significado semántico de los conceptos a evaluar, entre ellas podemos citar el análisis factorial, el análisis de cluster, etc...

Osgood (1952), en su libro *The Measurement of Meaning*, incluye los resultados de las primeras investigaciones factoriales que realizaron encontrando que, independientemente del concepto evaluado, había una serie de escalas que definían tres factores o dimensiones muy claras: *Valorativa, de Potencia y de Actividad*; se trata de variables «hipotéticas», en el contexto de los estudios de análisis factorial, pero se ha comprobado que tienen una correspondencia muy estrecha con las escalas semánticas reales definidas por los adjetivos: *Bueno - Malo, Fuerte - Débil* y *Activo - Pasivo*. Esta correspondencia, aunque estrecha, no es perfecta, por lo que se utilizará más de una escala bipolar para medir cada una de las dimensiones antes citadas. No hay un criterio estricto acerca del número de escalas que se debe utilizar para valorar de forma adecuada cada una de las dimensiones del espacio semántico; no obstante, se considera que unas seis escalas pueden ser suficientes.

A través de los distintos análisis factoriales realizados por Osgood y sus colaboradores se comprobó que en el primer factor, *el Valorativo* o *Evaluativo*, las escalas que obtenían una mayor saturación estaban formadas por adjetivos que implicaban una valoración del concepto:

Bueno – Malo, *Bonito – Feo,* *Sincero – Falso, etc.*

El segundo factor, *el de Potencia*, estaba formado por escalas cuyos adjetivos daban una idea de fuerza:

Fuerte – Débil, *Duro – Blando,* *Masculino – Femenino, etc.*

Por último, el factor correspondiente a la dimensión *de Actividad*, agrupaba aquellas escalas cuyos adjetivos denotaban un cierto sentido de movimiento:

Activo – Pasivo, *Rápido – Lento,* *Dinámico – Estático, etc.*

Después de analizar tanto la varianza total como la varianza común explicada por cada factor, se comprobó que el *factor evaluativo (valorativo)* era el que explicaba un mayor porcentaje de varianza.

Hasta ahora hemos hablado de tres factores o dimensiones subyacentes al espacio semántico; no obstante, se ha podido comprobar que, en función de la muestra utilizada y de los conceptos que se van a evaluar, pueden aparecer factores nuevos que habrá que ir identificando en cada caso, aunque los factores antes citados (valorativo, de potencia y de actividad) son los que tienen una mayor connotación semántica.

Podemos decir, por lo tanto, que un primer criterio para la selección de las escalas es su *composición factorial*, tratando de que cada una de las dimensiones o factores esté representada, al menos, por cuatro - seis escalas bipolares; estas escalas habrán de tener saturaciones muy altas en el factor que representan y bajas o nulas en el resto de los factores.

Otro criterio de selección es el *grado de relevancia* que tienen las distintas escalas para la evaluación de un determinado concepto. Puede suceder que escalas que tienen una alta saturación en uno de los factores, por ejemplo el valorativo, no tengan ninguna relevancia a la hora de evaluar un concepto. Por ejemplo, si tratáramos de juzgar una serie de fotografías para evaluar su composición estética, la escala valorativa *Bonita - Fea* puede ser muy relevante; sin embargo, la escala *Justo - Injusto*, que también es valorativa, puede no tener ninguna relevancia en nuestro trabajo. La inclusión de escalas poco relevantes, a lo único que conlleva es a una pérdida de información, ya que las respuestas emitidas por los sujetos, ante esas escalas, suelen situarse en el punto neutral.

No obstante, como ya plantearon Osgood y sus colaboradores (1976), hay veces que interesa utilizar deliberadamente escalas de este tipo; por ejemplo, en determinados estudios clínicos, cuando se quiere ver la influencia en la vida del paciente de determinadas personas, se pueden utilizar escalas como:

Caluroso – Frío, *Duro – Blando,* *Sabroso – Desabrido*

en lugar de

Apasionado – Frígido, *Agresivo – Tímido* y *Agradable – Desagradable.*

Otro criterio que se puede utilizar es el de la *estabilidad semántica* de la escala respecto a los conceptos y a los sujetos de una investigación. Si nos fijamos, por ejemplo, en la escala formada por los adjetivos: *Grande - Pequeño*, hemos de tener en cuenta que según sea el concepto que hemos de evaluar tendrá un significado u otro. Si utilizamos esta escala para evaluar conceptos tales como piedra, elefante, montaña, etc., tiene un uso denotativo; mientras que, si esa misma escala se utiliza para juzgar conceptos como: Dios, Patria o Presidente de Gobierno, puede ser usada de forma connotativa.

La elección de uno u otro criterio de selección dependerá, en último término, del tipo de investigación que se quiere llevar a cabo y del criterio del investigador.

4.4. Elaboración de la prueba piloto y aplicación

Una vez elaborada la lista de conceptos que se quieren evaluar, se puede pedir a una muestra de sujetos que califiquen cada uno de esos conceptos por medio de un adjetivo; de esta manera, podremos obtener una lista de adjetivos para calificar cada uno de los conceptos. A partir de esta lista, se puede hacer una selección previa de los adjetivos que se van a utilizar, siguiendo el criterio de máxima frecuencia utilizado ya por Osgood y sus colaboradores para la elaboración de su Diferencial Semántico, como hemos comentado anteriormente; es decir, se elegirán aquellos adjetivos que han aparecido con mayor frecuencia en la calificación de un concepto determinado; de esta manera, habremos hecho una preselección de adjetivos. El paso siguiente será buscar las palabras que tengan un significado opuesto al de cada uno de los adjetivos elegidos para formar las escalas bipolares.

Por último, se utilizará cualquiera de los criterios explicados en el apartado anterior para la selección de las escalas definitivas: criterio de la composición factorial de las escalas, el criterio de relevancia para la evaluación del concepto y, por último, el criterio de la estabilidad semántica respecto al concepto y a los sujetos.

Una vez seleccionados los conceptos y las escalas bipolares que se van a utilizar para evaluarlos, es necesario organizar unos y otras para su presentación y aplicación a una muestra de sujetos.

Aunque no hay una forma estándar de presentación del Diferencial Semántico, vamos a dar unas normas que creemos pueden facilitar la recogida de datos y su posterior análisis.

Por regla general, cuando la muestra que se va a utilizar no presenta problemas de alfabetización, se suele presentar el Diferencial Semántico en forma de cuadernillo, en el que la primera página está dedicada a las instrucciones de cumplimentación y en las páginas siguientes se recogen los distintos conceptos con sus escalas bipolares correspondientes; se suele utilizar una hoja para cada concepto.

Las instrucciones necesarias para la cumplimentación del Diferencial Semántico son muy sencillas y suelen ir orientadas a la naturaleza de la tarea, a la significación de las posiciones de las escalas y a la forma de marcarlas; asimismo, se ruega a los sujetos que respondan sin detenerse demasiado tiempo ya que la primera impresión es la que «vale».

Tampoco hay una norma generalizada acerca del número de conceptos y escalas que se deben aplicar de una sola vez; no obstante, es necesario que haya los suficientes como para cumplir con los objetivos de la investigación, pero se debe evitar el que sean tantos que produzcan aburrimiento y cansancio en los sujetos ya que, en este caso, sus respuestas no serían fiables.

Teniendo en cuenta que cada pareja de adjetivos que forman una escala bipolar representa una valoración positiva y negativa del concepto, y que estos adjetivos pueden estar situados aleatoriamente en un extremo o en el otro de la escala, es necesario que los valores numéricos asignados a las categorías de cada escala bipolar mantengan la dirección del continuo; así, el número más pequeño deberá corresponder a la categoría más próxima al adjetivo que representa una valoración negativa del concepto y el número más alto, corresponderá a la categoría más próxima al adjetivo que refleja una valoración positiva.

EJEMPLO:

AMOR

Bueno ---- 7 ----- 6 ----- 5 ----- 4 ------- 3 --------- 2 --------- 1 ----- Malo

o bien

----- 3 ----- 2 ----- 1 ----- 0 ----- (–1) ------ (–2) ----- (–3) ----

Pasivo ----- 1 ----- 2 ----- 3 ----- 4 ------- 5 --------- 6 --------- 7 ----- Activo

La escala numérica puede ser una de las dos que hemos mostrado en el ejemplo anterior (1 – 7), (–3 +3); también se podría haber elegido otra cualquiera, pero lo importante es que se mantenga la dirección del continuo, desde el punto más negativo, al más positivo. Obsérvese que en la segunda escala bipolar se ha cambiado la dirección de los valores numéricos.

La puntuación de cada sujeto en cada escala, es el valor numérico asignado a la categoría elegida; por lo tanto, los datos básicos, a partir de los cuales se realizarán todos los análisis necesarios, serán las puntuaciones obtenidas por los sujetos en cada una de las escalas que van a medir los distintos conceptos.

Se dispone de varios procedimientos y técnicas estadísticas para llevar a cabo el tratamiento y análisis de los datos; la utilización de uno u otra va a depender del objetivo de la investigación que

se lleve a cabo. Unas veces nos interesarán los datos grupales y otras serán los datos individuales los que merezcan nuestra atención.

En cualquier caso podremos obtener las siguientes:

Medidas descriptivas:

a) Puntuaciones escalares

— Se puede averiguar la puntuación de un sujeto o grupo de sujetos en cada una de las escalas.

— Averiguar la puntuación media obtenida por la muestra de sujetos en cada una de las escalas bipolares utilizadas para evaluar cada concepto.

— La puntuación media de todas las escalas que evalúan un concepto determinado, tanto a nivel individual como grupal.

— La puntuación media de todos los conceptos y de todos los sujetos, etc.

b) Puntuaciones factoriales

Las puntuaciones factoriales se obtienen con el fin de averiguar la puntuación que corresponde a cada una de las dimensiones subyacentes o factores. Se pueden obtener tanto a nivel individual como a nivel grupal. Cada una de las puntuaciones factoriales representa la reacción afectiva de un sujeto, o grupo de sujetos, a un concepto determinado en una de las dimensiones del Diferencial Semántico. Para su obtención, se calcula la media de las puntuaciones escalares que definen cada una de las dimensiones o factores.

Si, por ejemplo, queremos hallar la puntuación factorial que ha obtenido una muestra de sujetos en la dimensión *Actividad* para el concepto MATRIMONIO, y contamos con tres escalas bipolares para definir esta dimensión, el primer paso será averiguar la puntuación media de cada una de las escalas y, posteriormente, hallar la media de estos valores que corresponderá a la puntuación factorial del grupo, en la dimensión estudiada, y para el concepto «Matrimonio». Si, por el contrario fuera la puntuación factorial de un único sujeto la que quisiéramos obtener, bastaría hallar la media de las puntuaciones obtenidas por el sujeto en las escalas que definen la dimensión «Actividad».

Supongamos que las tres escalas utilizadas, así como el número de sujetos de la muestra que respondieron en cada una de las categorías de las mismas, son las que ofrecemos a continuación:

MATRIMONIO

	Número de sujetos en cada categoría	
Pasivo	----- 4 ------- 6 ------- 8 ------ 10 ----- 12 ----- 30 ----- 50 -----	**Activo**
Escala	1 2 3 4 5 6 7	
	Número de sujetos en cada categoría	
Lento	----- 6 ------- 4 ------ 10 ------ 8 ------ 50 ----- 30 ----- 12 -----	**Rápido**
Escala	1 2 3 4 5 6 7	
	Número de sujetos en cada categoría	
Estático	----- 10 ----- 30 ----- 50 ----- 12 ------ 4 ------ 6 ------- 8 ------	**Dinámico**
Escala	7 6 5 4 3 2 1	

Los números que aparecen en la parte superior, corresponden a los sujetos de la muestra que clasificaron el concepto MATRIMONIO en una categoría determinada de la escala bipolar correspondiente. Así por ejemplo, hay 10 sujetos que, en la escala bipolar *Estático-Dinámico*, asignaron un 7 al concepto MATRIMONIO, 50 sujetos le asignaron un 5, y 8 sujetos le asignaron un 1. La media de cada escala es:

Media de Pasivo – Activo = 5,58

Media de Lento – Rápido = 4,92

Media de Estático – Dinámico = 4,83

Para obtener estas puntuaciones medias se aplica la fórmula de la media, multiplicando el número de sujetos que hay en cada categoría por el valor numérico de dicha categoría y dividiendo por el número total de sujetos. Así por ejemplo, la media de la escala PASIVO – ACTIVO se calcularía de la siguiente manera:

$$\overline{X} = \frac{\Sigma f \cdot X}{N} = \frac{(4\times 1)+(6\times 2)+(8\times 3)+(10\times 4)+(12\times 5)+(30\times 6)+(50\times 7)}{120} = \frac{670}{120} = 5{,}58$$

De la misma forma se irían calculando las medias del resto de las escalas.

A partir de estas puntuaciones medias, calculamos la puntuación factorial del grupo en la dimensión *Actividad*, para el concepto MATRIMONIO. Para ello basta sumar las puntuaciones obtenidas y dividir por el número de escalas, en nuestro caso 3.

$$PF = \frac{5{,}58 + 4{,}92 + 4{,}83}{3} = 5{,}11$$

Teniendo en cuenta que la escala utilizada tiene el punto neutral en el valor numérico 4, una puntuación factorial de 5,11 puntos indicará que la muestra considera el concepto MATRIMONIO ligeramente activo.

En este ejemplo sólo hemos averiguado la puntuación factorial para la dimensión *Actividad*; el mismo procedimiento habría que seguir para averiguar las puntuaciones factoriales correspondientes a las demás dimensiones.

5. LA TÉCNICA DE GUTTMAN

La técnica presentada por Louis Guttman se desarrolló como un modelo alternativo a las técnicas elaboradas por Thurstone y Likert para la medida de las actitudes, aunque también puede ser utilizado para la construcción de tests en los que haya respuestas correctas o incorrectas.

Se le conoce, generalmente, como Escalograma de Guttman y está diseñado de tal forma que es posible conseguir la ordenación tanto de los sujetos como de los estímulos (los ítems) respecto a una dimensión determinada, asegurando, si los resultados demuestran que los datos se ajustan al modelo propuesto por Guttman, que el conjunto de ítems que conforman la escala miden una única dimensión y que, por lo tanto, la puntuación total que se asigne a los sujetos tenga significado psicológico y pueda ser interpretable, cosa que no ocurriría en el caso de que los ítems hicieran referencia a más de una dimensión.

En este sentido la técnica del escalograma de Guttman está más orientada a probar la existencia de una única dimensión subyacente al conjunto de ítems de la escala (su unidimensionalidad) que al proceso de construcción de la misma.

El modelo está basado en la idea de que es posible ordenar los estímulos de manera que si un sujeto responde correctamente (o favorablemente) a un estímulo concreto, lo hará también a todos los que estén situados por debajo de dicho estímulo en la escala establecida, y si un sujeto no responde correctamente (o favorablemente) ante un determinado estímulo, tampoco lo hará al contestar a los que estén situados por encima de él en la escala.

De esta manera, tanto los sujetos como los estímulos pueden representarse, a lo largo de un continuo, formando una escala denominada *escala de entrelazamiento*. En esta escala, cada sujeto estará situado entre dos estímulos y su orden será el inmediato superior al del último estímulo que ha contestado correctamente (o favorablemente) y el inmediato inferior al del primer estímulo que no contesta correctamente (o favorablemente).

Si suponemos que los datos que se presentan a continuación son las respuestas de cinco sujetos a cuatro elementos dicotómicos y que el 1 significa que los sujetos han mostrado su acuerdo (o acertado) con el elemento y el cero que han mostrado su desacuerdo (o fallado), estaríamos ante una escala de Guttman perfecta.

El primer sujeto ha acertado (ha respondido favorablemente) todos los elementos de la prueba, por lo tanto en la escala ocupará el puesto quinto ya que dejará por debajo de sí a los cuatro elementos; sin embargo, el sujeto 5, que no ha acertado ningún elemento, estará situado por debajo de todos los ellos (Tabla 3.5).

Cuando tenemos una escala perfecta, la organización de los datos dará lugar a una matriz triangular y, a partir de la puntuación que tenga cada sujeto en la escala, se podrá predecir con exactitud cuál ha sido la respuesta dada ante cada uno de los elementos. Así una puntuación de 3 en la escala perfecta, indicaría que el sujeto ha respondido correctamente a tres de los cuatro elementos, pero no a tres cualesquiera, sino a los tres primeros; es decir, su patrón de respuestas habrá sido: (1 1 1 0).

TABLA 3.5
Matriz de datos

	Elementos				
Sujetos	A	B	C	D	**Puntuación Sujeto**
1	1	1	1	1	4
2	1	1	1	0	3
3	1	1	0	0	2
4	1	0	0	0	1
5	0	0	0	0	0
Puntos Ítem	4	3	2	1	

La representación de la escala de entrelazamiento sería:

5 A 4 B 3 C 2 D 1

En la práctica es casi imposible obtener escalas perfectas, el problema consiste en determinar qué grado de desviación, respecto a la escala perfecta, se debe tolerar para aceptar que los datos obtenidos se ajustan al modelo de Guttman.

El interés de esta técnica de escalamiento está, como ya hemos mencionado, en que trata de medir conceptos unidimensionales a través de todo un universo de atributos. Para llevar a cabo el análisis del escalograma es necesario seguir una serie de pasos que vamos a resumir:

a) Establecer la forma de evaluar el error o desviación de los datos respecto al modelo.

b) Ordenar los datos de manera que se ajusten lo más posible a una escala perfecta.

c) Evaluar el grado de aproximación de los datos obtenidos a la escala perfecta.

d) Establecer reglas para fijar la posición, en el continuo, de aquellos sujetos cuyo patrón de respuestas se separa del patrón ideal.

e) Establecer reglas o normas para comprobar si el conjunto de elementos elegido, es escalable según el modelo de Guttman.

Para la elaboración de escalas de Guttman, los ítems se eligen sobre la base de que miden una determinada actitud que es la que se quiere medir, y se van eligiendo de manera que su grado de extremosidad respecto a la variable medida vaya en aumento; es decir, en primer lugar se seleccionan unos pocos ítems con los cuales, a priori, se muestren de acuerdo la mayoría de las personas; una vez seleccionados estos ítems, se seleccionarán otros cuantos cuya aceptación indique una actitud moderada hacia el objeto de estudio y, finalmente, se incluirán ítems cuyo contenido refleje una actitud extrema. De esta manera, si la escala es correcta, los sujetos que respondan favorablemente a los ítems más extremos deberán hacerlo en el mismo sentido a todos los ítems que representen actitudes menos extremas.

Nota: Si se tratara de elaborar una escala de conocimientos en la que hubiera respuestas correctas o incorrectas, los ítems se ordenarían en función de su grado de dificultad, desde los más fáciles a los más difíciles.

5.1. Evaluación del error en el modelo

En primer lugar es necesario definir lo que se entiende por error en este contexto; llamaremos error a la desviación del patrón de respuestas observado respecto al patrón de respuestas ideal requerido por el modelo.

Aunque hay varios procedimientos para averiguar el número de errores o desviaciones, vamos a fijarnos en el propuesto por Goodenough (1944) y Edwards (1948), que está basado en el número de desviaciones encontradas en la escala empírica respecto a la escala ideal.

Supongamos 4 sujetos que han respondido a cuatro elementos y que sus patrones de respuesta han sido los que figuran en la tabla siguiente, tabla 3.6, junto con el patrón ideal y el número de errores.

¿Cómo se ha llevado a cabo el recuento de errores?

Si tomamos como ejemplo el patrón de respuestas correspondiente al sujeto D (- + - -) el número de errores sería 2 pues tendríamos que hacer dos cambios para obtener el patrón ideal (el elemento 2 debería estar contestado desfavorablemente y el 1 favorablemente). En el patrón de respuestas del sujeto A no hay ningún error ya que coincide con el patrón de respuestas ideal.

TABLA 3.6
Matriz de datos para el recuento de errores

Sujetos	Patrón respuestas	Patrón Ideal	Errores
A	+ + + +	+ + + +	0
B	− + + +	+ + + −	2
C	− − + +	+ + − −	4
D	− + − −	+ − − −	2

5.2. Pasos a seguir para la elaboración de una escala

Una vez que se cuenta con el conjunto de ítems es necesario aplicarlos a una muestra de sujetos.

Vamos a explicar el proceso utilizando ítems dicotómicos que son aquellos que no admiten más de dos opciones de respuesta: favorable (1) – desfavorable (0), acierto (1) – fallo (0), etc.

El hablar de respuestas favorables o correctas dependerá del contexto en el que se trabaje, como ya indicamos anteriormente; si se trata de una escala de actitudes, un 1 podría significar una actitud favorable y un 0 desfavorable; si se trata de una prueba de aptitud el 1 podría significar una respuesta correcta o acierto y el 0 una respuesta incorrecta o fallo.

Si tenemos un conjunto de ítems dicotómicos el número de patrones de respuesta distintos que se podrían producir sería 2^n, siendo n el número de ítems de la escala. Si tenemos 3 elementos dicotómicos, el número posible de patrones de respuesta que se podrían producir sería 2^3, igual a 8; ahora bien, de esos 8 posibles patrones de respuesta, sólo hay cuatro que se ajustan al modelo de Guttman: (1 1 1), (1 1 0), (1 0 0) y (0 0 0).

Supongamos la siguiente escala:

A | 1 B | 2 C | 3 D |

el sujeto D ha contestado correctamente (o favorablemente) a todos los elementos, por lo que se situará por encima de todos ellos en la escala, su patrón de respuestas habría sido (1 1 1), el sujeto A, por el contrario, no ha contestado correctamente a ninguno de los elementos de la escala, por lo tanto se colocará por debajo de todos ellos en el continuo y su patrón de respuestas habrá sido (0 0 0). El sujeto B ha respondido favorablemente al ítem 1 y su patrón de respuestas es (1 0

0) y el sujeto C ha respondido favorablemente a los ítems 1 y 2 luego su patrón de respuestas es (1 1 0).

Cuando el número de sujetos y/o de estímulos es muy grande, esto es muy laborioso; el procedimiento más sencillo es ordenar los datos en una matriz que tenga por columnas a los estímulos y por filas a los sujetos (o a la inversa), de manera que en cada casilla aparezca la respuesta que cada sujeto emite ante cada estímulo y, así poder elaborar la escala.

Los pasos a seguir para la elaboración de la escala son:

a) Averiguar para cada uno de los sujetos el número de respuestas favorables o correctas según su patrón de respuestas (puntuación del sujeto).

b) Averiguar la proporción de sujetos que responden correctamente (o favorablemente) a cada ítem.

c) En función de los resultados obtenidos en el punto anterior, intercambiar las columnas correspondientes a los ítems de manera que estos queden ordenados en función de la mayor o menor proporción de respuestas favorables o correctas.

d) Reordenar las filas correspondientes a los sujetos de forma que éstos queden ordenados desde el que obtuvo una mayor puntuación (mayor número de respuestas favorables o correctas) hasta el que la obtuvo más pequeña.

Después de haber realizado todos estos pasos, si la matriz de datos se ajustara a una escala acumulativa perfecta, el resultado sería una matriz triangular; en caso contrario, deberíamos hacer el recuento de los errores para comprobar el grado de ajuste de los datos obtenidos al modelo de Guttman.

EJEMPLO:

En una escala de actitudes un grupo de 10 sujetos obtuvo los siguientes patrones de respuesta ante 8 elementos.

Como puede apreciarse en la tabla 3.7, se ha calculado la puntuación total de los sujetos y el número de los mismos que respondieron correctamente a cada uno de los elementos.

TABLA 3.7
Matriz de respuestas

	Elementos								
Sujetos	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Puntuación Sujeto**
A	1	1	1	1	1	1	1	1	8
B	0	0	0	0	0	0	0	0	0
C	1	1	1	0	1	1	0	0	5
D	1	0	0	0	1	0	0	0	2
E	1	1	1	1	1	1	1	1	8
F	1	1	1	0	0	0	0	0	3
G	1	1	1	1	1	0	1	0	6
H	0	0	0	1	0	0	0	0	1
I	1	0	0	0	0	0	0	0	1
J	1	1	1	1	0	0	1	1	6
Aciertos	8	6	6	5	5	3	4	3	

El paso que hay que dar ahora es ordenar las columnas desde el elemento más difícil al más fácil (desde el menos acertado al más acertado), si se trata de una escala de actitud en la que no se puede hablar de elementos más difíciles o fáciles, la ordenación se haría desde el menos aceptado al más aceptado.

Dado que hay una serie de elementos que tienen el mismo grado de dificultad o de aceptación (según el contexto), de momento es indiferente el orden en que se sitúen, esto pasa con el elemento 6 y 8, el 5 y 4, el 2 y 3.

TABLA 3.8
Matriz de columnas ordenadas

	Elementos								
Sujetos	**6**	**8**	**7**	**4**	**5**	**2**	**3**	**1**	**Puntuación Sujeto**
A	1	1	1	1	1	1	1	1	8
B	0	0	0	0	0	0	0	0	0
C	1	0	0	0	1	1	1	1	5
D	0	0	0	0	1	0	0	1	2
E	1	1	1	1	1	1	1	1	8
F	0	0	0	0	0	1	1	1	3
G	0	0	1	1	1	1	1	1	6
H	0	0	0	1	0	0	0	0	1
I	0	0	0	0	0	0	0	1	1
J	0	1	1	1	0	1	1	1	6
Aciertos	3	3	4	5	5	6	6	8	

Una vez ordenadas las columnas se ordenan las filas; podemos situar en primer lugar el sujeto que obtuvo una menor o una mayor puntuación y, posteriormente, continuar en el orden iniciado.

Como puede apreciarse en la tabla 3.9, aunque no hemos obtenido una matriz triangular perfecta, ya que aparecen algunos patrones de respuesta que no se ajustan al patrón ideal, sin embargo una vez que se ha realizado la ordenación de filas y columnas se observa un cierto parecido a lo que podría ser una matriz triangular.

Una vez ordenadas las filas y las columnas de la matriz, se realiza el recuento de errores para ver la bondad de ajuste de los datos al modelo. Es el momento de analizar si los ítems que han recibido el mismo número de aceptaciones deben quedar como están en la ordenación final o, por el contrario, se debe invertir su orden para obtener un menor número de errores; en nuestro caso la inversión del orden de las columnas correspondientes a estos ítems no disminuye el número de errores.

Si nuestros datos se ajustaran perfectamente, la matriz resultante hubiera sido una matriz triangular, en la medida en que esto no se verifique será que se han cometido una serie de errores o desviaciones respecto al modelo Supongamos el sujeto J que ha obtenido una puntuación de 6 puntos, si su patrón de respuestas se ajustara al modelo de Guttman los elementos que debería haber acertado serían los seis más fáciles, en nuestro caso todos menos el 6 y el 8; sin embargo, vemos que ha fallado el elemento 5 y ha acertado el elemento 8, hay dos desviaciones respecto al patrón de respuestas ideal. En cambio si observamos el sujeto G, vemos que su patrón de respuestas se ajusta exactamente al patrón de respuestas ideal, ya que obtiene una puntuación de 6 y los elementos a los cuales contesta correctamente son los seis más fáciles.

TABLA 3.9
Matriz ordenada de filas y columnas

	Elementos									
Sujetos	**6**	**8**	**7**	**4**	**5**	**2**	**3**	**1**	**Puntuación Sujeto**	**Número errores**
B	0	0	0	0	0	0	0	0	0	0
H	0	0	0	1	0	0	0	0	1	2
I	0	0	0	0	0	0	0	1	1	0
D	0	0	0	0	1	0	0	1	2	2
F	0	0	0	0	0	1	1	1	3	0
C	1	0	0	0	1	1	1	1	5	2
G	0	0	1	1	1	1	1	1	6	0
J	0	1	1	1	0	1	1	1	6	2
E	1	1	1	1	1	1	1	1	8	0
A	1	1	1	1	1	1	1	1	8	0
Aciertos	3	3	4	5	5	6	6	8		

De esta manera iremos analizando los patrones de respuesta de todos los sujetos de la muestra y contando el número de errores.

Guttman propuso que como criterio de bondad de ajuste se utilizara el Coeficiente de Reproductividad (C.R.), cuya fórmula viene expresada de la siguiente manera:

$$C.R. = 1 - \frac{\text{Número de errores}}{\text{Número total de respuestas}} = 1 - \frac{\text{Número de errores}}{(\text{Número de ítems} \times \text{Número de sujetos})}$$

Según el criterio establecido por Guttman, diremos que unos datos empíricos se ajustan al modelo de Guttman si su coeficiente de reproductividad es igual o mayor que 0,90.

En nuestro ejemplo el coeficiente de reproductividad será:

$$C.R. = 1 - \frac{8}{10 \times 8} = 1 - 0{,}10 = 0{,}90$$

luego podemos decir que, aunque en el límite, nuestros datos son escalables según el modelo de Guttman.

6. DIFERENCIAS ENTRE LAS DISTINTAS TÉCNICAS

Las escalas elaboradas mediante el Diferencial Semántico se diferencian de las otras tres (Thurstone, Likert y Guttman) fundamentalmente en el formato de los ítems que presentan. En un Diferencial Semántico los ítems consisten en una serie de conceptos que han de ser evaluados por los sujetos mediante sus respuestas a una serie de escalas ancladas por adjetivos bipolares. En las otras escalas, los ítems están formados por enunciados o frases, no adjetivos.

Las diferencias entre las escalas de Thurstone, Likert y Guttman pueden analizarse considerando que la actitud existe a lo largo de un continuo subyacente, que el punto medio del continuo indica un cambio en la dirección de la actitud y que la distancia desde el punto medio en una u otra dirección indica la intensidad de la misma. Partiendo de esta consideración, la colocación de los ítems a lo largo del continuo diferencia las escalas de Likert de las de Thurstone y Guttman.

En las escalas de Likert, los enunciados de los ítems se sitúan sólo (o muy cerca) en los dos extremos del continuo, deberán indicar una actitud positiva o negativa. En efecto, se excluyen los ítems cuyos enunciados puedan ser interpretados como representantes de los puntos del continuo situados alrededor del punto medio. Por el contrario, en las escalas de Thurstone y de Guttman, es necesario incluir ítems que cubran todo el continuo de la actitud, desde uno de los extremos hasta el otro.

Las escalas de Guttman son acumulativas, esto las diferencia de las de Thurstone. Una respuesta positiva o favorable a un ítem situado en un punto del continuo de actitud, implica una respuesta positiva a todos los ítems que están situados en el continuo a la izquierda del ítem en cuestión. Las escalas de Thurstone no son acumulativas. Aunque los ítems se deben redactar de manera que reflejen sentimientos separados en intervalos aparentemente iguales a lo largo del continuo, no hay que asumir que las respuestas sean acumulativas. El supuesto que se asume es que las respuestas positivas a los ítems, dadas por un sujeto, deben estar reunidas todas alrededor de un punto concreto del continuo; no es lógico pensar que un sujeto que esté de acuerdo con una serie de ítems que demuestran una actitud muy favorable (valores escalares alrededor del punto 9, por ejemplo) elija también ítems cuyos valores escalares se sitúen en el polo opuesto.

Desde el punto de vista de la medición la técnica de Thurstone asume un nivel de medida de intervalos, aunque esto esté hoy día bastante cuestionado, mientras que las otras técnicas dan lugar a escalas ordinales.

La selección de los ítems definitivos de la escala de Thurstone se basa en los valores asignados por los jueces a cada uno de los ítems. En el resto de las escalas es necesario aplicar la escala a una muestra representativa de sujetos.

De las cuatro técnicas explicadas, las escalas de Likert son las más comúnmente utilizadas ya que son rápidas de administrar y puntuar, se adaptan fácilmente para poder medir la mayoría de las actitudes y si están bien construidas proporcionan información fiable. Sin embargo tienen también algunos inconvenientes entre los que merecen destacar, por una parte, la facilidad que tienen los sujetos para emitir respuestas falsas, socialmente deseables, en lugar de manifestar su actitud real ante el enunciado de cada ítem y, por otra, el que los intervalos entre los puntos de la escala no representan cambios iguales en la actitud medida en todos los sujetos (Keeves, 1988).

7. EJERCICIOS DE AUTOEVALUACIÓN

1. En un estudio realizado por la empresa consultora Wilkingston se pretendió estudiar las valoraciones que los españoles hacen de cuatro líderes políticos: A B C y D y si existen diferencias en cuanto al género en estas valoraciones. La matriz 1, representa las respuestas de los varones y la matriz 2 la de las mujeres. Los elementos de cada una de las matrices representan el número de sujetos que valoran más positivamente al político representado en la columna que al de la fila.

Elaborar la escala correspondiente a los varones y a las mujeres e indicar cuál de los dos grupos es más homogéneo respecto a su actitud:

Matriz 1 (Varones)

	A	B	C	D
A	—	1.500	1.000	200
B	2.500	—	1.900	500
C	3.000	2.100	—	1.000
D	3.800	3.500	3.000	—
Σ	9.300	7.100	5.900	1.700

Matriz 2 (Mujeres)

	A	B	C	D
A	—	1.000	500	1.000
B	3.000	—	3.000	2.500
C	3.500	1.000	—	500
D	3.000	1.500	3.500	—
Σ	9.500	3.500	7.000	4.000

2. Se quiere llevar a cabo un estudio acerca de las campañas publicitarias de tres marcas de detergentes. Para ello, se elige una muestra de amas de casa y, a cada una de ellas, se les pide que asignen, cada una de las marcas de detergente cuya campaña se va a estudiar, a una serie de 9 categorías ordenadas. La asignación habrá de hacerse en función del grado en que la campaña publicitaria define el producto que representa, no en función de si les gusta o no. Deberán asignar a la categoría **A** la marca de detergente cuya campaña les parezca peor y a la categoría **I** la que les parezca mejor. Los resultados se recogen en la matriz adjunta:

Productos	A	B	C	D	E	F	G	H	I
1	10	20	30	40	50	30	20	15	5
2	5	10	15	50	30	40	20	30	20
3	50	40	30	30	20	20	15	10	5

2.1. Averiguar el valor escalar de cada producto

3. Se quiere elaborar una escala tipo Likert para evaluar la actitud de los alumnos de Psicología ante el nuevo plan de estudios. Para ello, se ha elaborado una prueba piloto tipo Likert formada por 5 ítems a los que se han asignado valores de 1-5, correspondiendo el valor 5 a una actitud más positiva hacia el plan de estudios, y se ha aplicado a una muestra de 12 sujetos (se trata de un ejemplo).

Los resultados obtenidos fueron los siguientes:

		SUJETOS											
		1	2	3	4	5	6	7	8	9	10	11	12
ÍTEMS	A	3	4	5	2	3	5	4	3	4	2	2	1
	B	1	3	2	4	3	2	4	3	5	1	2	3
	C	2	1	3	3	4	1	5	1	4	5	3	4
	D	5	5	2	1	1	4	4	3	5	2	4	3
	E	4	5	4	2	3	1	3	3	4	4	4	5

3.1. Si la puntuación más alta, corresponde a una actitud más favorable, ¿qué sujeto manifiesta una actitud más desfavorable?

3.2. ¿Podemos decir que el grupo es bastante homogéneo respecto a la actitud que manifiestan?

4. Se quiere llevar a cabo una investigación transcultural para ver las diferencias de significado de una serie de conceptos. Para ello, se han utilizado dos muestras de distintas culturas, a las que se les ha aplicado el siguiente Diferencial Semántico:

CONCEPTOS	ESCALAS BIPOLARES	
GUERRA AMOR	Sucio	Limpio (E)
	Activo	Pasivo (A)
	Grande	Pequeño (P)
	Valioso	Despreciable (E)
	Caliente	Frío(A)
	Fuerte	Débil (P)
	Agradable	Desagradable (E)
	Profundo	Superficial (P)
	Rápido	Lento (A)

Cada uno de los conceptos que se presentaron anteriormente, se evaluó a través de esas nueve escalas. Previamente se había comprobado que a dichas escalas subyacían las tres di-

mensiones de la trilogía clásica E.P.A.: Evaluativa, Potencia y Actividad. Entre paréntesis aparece la dimensión en la cual satura cada una de las escalas. Las medias obtenidas por cada uno de los grupos, en cada una de las escalas bipolares utilizadas para evaluar los distintos conceptos, son las que aparecen a continuación.

La escala numérica asignada a cada par de adjetivos bipolares fue de (-3 a +3) y el punto neutral de la escala el 0. Averiguar las puntuaciones factoriales para cada concepto, en cada uno de los grupos.

GUERRA

	Grupo A	Grupo B
Sucio - Limpio	-3	-2
Activo - Pasivo	3	3
Grande - Pequeño	3	2
Valioso - Despreciable	-2	-1
Caliente - Frío	1	2
Fuerte - Débil	2	3
Agradable - Desagradable	-3	-2
Profundo - Superficial	3	3
Rápido - Lento	-2	3

AMOR

	Grupo A	Grupo B
Sucio - Limpio	3	3
Activo - Pasivo	2	1
Grande - Pequeño	3	2
Valioso - Despreciable	3	3
Caliente - Frío	2	2
Fuerte - Débil	2	1
Agradable - Desagradable	3	3
Profundo - Superficial	2	2
Rápido - Lento	2	2

5. Ejercicios conceptuales

A continuación se presentan una serie de afirmaciones que deberán leerse atentamente y responder si son correctas o incorrectas.

1. Según Thurstone, cuando a un sujeto (o grupo de sujetos) se le presenta un estímulo para que emita un juicio acerca de él, se produce en el sujeto un proceso discriminante.
2. Un mismo estímulo suscita siempre en el sujeto (o sujetos) el mismo proceso discriminante.
3. Según el modelo de Thurstone, si un estímulo suscita en el sujeto una gran ambigüedad a la hora de asignarle un valor en el continuo psicológico, la desviación típica de la distribución discriminante será pequeña.
4. En el modelo de Thurstone, el valor escalar de un estímulo es la media de los valores asignados por el sujeto (o sujetos), a dicho estímulo, a través de los distintos procesos discriminantes.
5. La distribución de los valores asignados por el sujeto a cada uno de los estímulos, a través de los distintos procesos discriminantes, es una distribución normal.
6. En el modelo de escalamiento de Thurstone, los sujetos actúan como instrumentos de medida.
7. En el método de las comparaciones binarias los sujetos asignan, de forma directa, el valor en el continuo psicológico a cada uno de los estímulos.
8. Las escalas derivadas de la aplicación del modelo escalar de Thurstone se incluyen dentro del grupo conocido por el nombre de «escalas de Juicio».
9. Si a un sujeto se le presentan varias veces una serie de estímulos, para que les asigne un valor en el continuo psicológico, cada uno de ellos dará lugar a una distribución discriminativa distinta.
10. Si un estímulo *(K)* es preferido a otro *(J)* por el sujeto, el valor escalar de *(K)* será mayor que el de *(J)*.
11. Las escalas de Likert se utilizan para escalar estímulos.
12. Para la elaboración de una escala de Likert se utiliza la prueba de jueces.
13. El Diferencial Semántico se utiliza para medir el significado afectivo de los conceptos.
14. Para evaluar los distintos conceptos en el Diferencial Semántico, se utilizan escalas bipolares.
15. Dado un concepto cualquiera, podemos decir que su significado semántico vendrá definido, exclusivamente, por las dimensiones: Evaluativa, Potencia y Actividad.
16. Para evaluar el significado de un concepto, basta utilizar una escala bipolar por cada una de las dimensiones subyacentes al mismo.
17. La técnica de Guttman da lugar a escalas de entrelazamiento.
18. Si el ajuste de los datos al modelo de Guttman fuera perfecto la matriz resultante sería triangular.

19. El coeficiente de reproductividad puede ser negativo.
20. Se consideran errores en una escala de Guttman, a las desviaciones encontradas en el patrón de respuestas de los sujetos respecto al patrón que deberían haber obtenido si el ajuste de los datos al modelo fuera perfecto.

8. SOLUCIONES A LOS EJERCICIOS DE AUTOEVALUACIÓN

1. Las matrices de frecuencias se transforman en proporciones, redondeadas a dos decimales:

Matriz 1

	A	B	C	D
A	—	0,38	0,25	0,05
B	0,62	—	0,48	0,12
C	0,75	0,52	—	0,25
D	0,95	0,88	0,75	—
Σ	2,32	1,78	1,48	0,42

Matriz 2

	A	B	C	D
A	—	0,25	0,12	0,25
B	0,75	—	0,75	0,62
C	0,88	0,25	—	0,12
D	0,75	0,38	0,88	—
Σ	2,38	0,88	1,75	0,99

Se observa que así como en la primera matriz están ordenados los estímulos en función de las preferencias de los varones, en la segunda matriz es necesario llevar a cabo una ordenación antes de transformar la matriz de proporciones a puntuaciones típicas. Para mantener la misma dirección en la ordenación de las dos matrices es necesario situar el estímulo B en el último puesto. La ordenación quedaría así: A C D B:

Matriz 2

	A	C	D	B
A	—	0,12	0,25	0,25
C	0,88	—	0,12	0,25
D	0,75	0,88	—	0,38
B	0,75	0,75	0,62	—
Σ	2,38	1,75	0,99	0,88

Compruébese como la suma de los elementos simétricos de la matriz suman la unidad.

Una vez obtenidas las matrices de proporciones se transforman en matrices de puntuaciones típicas, utilizando para ello, la tabla de las áreas bajo la curva normal de probabilidad.

Matriz 1

	A	B	C	D
A	0,00	-0,31	-0,67	-1,64
B	0,31	0,00	-0,05	-1,18
C	0,67	0,05	0,00	-0,67
D	1,64	1,18	0,67	0,00
Σ =	2,62	0,92	-0,05	-3,49
Σ/n =	0,66	0,23	-0,012	-0,87

Matriz 2

	A	C	D	B
A	0,00	-1,18	-0,67	-0,67
C	1,18	0,00	-1,18	-0,67
D	0,67	1,18	0,00	-0,31
B	0,67	0,67	0,31	0,00
Σ =	2,52	0,67	-1,54	-1,65
Σ/n =	0,63	0,17	-0,38	-0,41

La suma de los elementos simétricos de estas matrices es igual a cero.

A partir de la matriz de puntuaciones típicas, la mejor estimación que podemos hacer de los valores escalares de los estímulos es la media de su columna correspondiente, por eso se han sumado las columnas y, el resultado se ha dividido por 4 que es el número de estímulos que teníamos.

En la última fila aparecen los valores escalares correspondientes a cada uno de los estímulos; como ya comentamos anteriormente, estas escalas tenían el inconveniente de los valores negativos, ya que la suma de todos ellos ha se ser cero (excepto por problemas de redondeo); por eso, se puede hacer una transformación lineal y situar el origen de la escala en el valor más pequeño. En el caso de la escala correspondiente a los varones el valor más pequeño es (-0,87) que corresponde al político D, a ese estímulo le hacemos corresponder el cero de la escala; como lo que se ha hecho ha sido añadir al valor escalar que tenía una constante igual pero de signo contrario (0,87), esa constante habrá que sumársela al resto de los valores escalares para que la distancia que había entre ellos no se modifique por el cambio de origen. La escala resultante para la muestra de varones será:

D.....C.....B.....A.....

0 0,86 1,10 1,53

Llevados los valores sobre una escala de intervalos, las distancias entre los distintos políticos en la escala de preferencias de los españoles sería aproximadamente como la que aparece en la representación.

El político menos valorado es el D y por el que muestran sus preferencias es el A.

En la muestra de mujeres el valor escalar más bajo corresponde al político B (-0,41), si asignamos a este estímulo el cero de la escala, deberemos sumar al resto de los valores escalares una constante igual a 0,41 para obtener los valores transformados; así, la escala resultante para la muestra de mujeres será:

B....D.....C.....A

0 0,03 0,58 1,04

En la escala de mujeres se observa que hay dos políticos que son valorados negativamente por las mujeres y que están muy cerca en la escala uno de otro (sus valores escalares son casi iguales).Tanto los hombres como las mujeres están de acuerdo en cuanto a sus preferencias por el político A.

Un índice del grado de homogeneidad de la muestra respecto a los estímulos analizados es el recorrido de la escala, entendiendo por recorrido la distancia que hay desde el origen al valor escalar más alto; en la medida en que el recorrido es mayor, hay una mayor homogeneidad.

Parece, por lo tanto, que la actitud de los varones respecto a los cuatro políticos objeto de estudio es más homogénea que la de las mujeres, ya que el recorrido de la escala de los varones es mayor. Hay un mayor acuerdo entre ellos en cuanto a sus preferencias por los políticos.

En la escala de la muestra de mujeres el recorrido es muy pequeño y, sobre todo respecto a los políticos D y B, no parece haber habido una actitud uniforme en cuanto a sus preferencias ya que están prácticamente juntos en el continuo psicológico.

2. Para calcular el valor escalar de cada ítem es necesario calcular la mediana.

Se asume que el método que se va a utilizar es el de los intervalos aparentemente iguales. A cada categoría se le asigna un valor numérico de la siguiente manera. Un 1 a la categoría A, un 2 a la B, y así sucesivamente hasta asignar el valor 9 a la categoría I.

Se van calculando las frecuencias acumuladas de manera que vemos que hay 220 amas de casa que actúan como *jueces*. Para cada producto habrá que calcular el valor de la mediana.

Vamos a ir comentándolo paso a paso. Sustituyendo las letras correspondientes a las categorías por sus valores numéricos.

	Categorías								
Producto 1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Frecuencias (acumuladas)	10	30	60	100	150	180	200	215	220

el 50% de 220 amas de casa son 110. Esa frecuencia se encuentra en la categoría E que tiene un valor numérico de 5 y cuyos límites son 4,5 – 5,5.

$$Md. = 4,5 + \frac{1}{50}(110 - 100) = 4,5 + 0,20 = 4,70$$

	Categorías								
Producto 2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Frecuencias (acumuladas)	5	15	30	80	110	150	170	200	220

$$Md. = 4,5 + \frac{1}{30}(110 - 80) = 5,50$$

al mismo resultado se hubiera llegado sin necesidad de haber aplicado la prueba. Hubiera bastado con tener en cuenta que el límite superior de la categoría E, cuyo valor es 5,5, deja por debajo a 110 amas de casa (el 50%) que evaluaron el producto.

	Categorías								
Producto 3	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Frecuencias (acumuladas)	50	90	120	150	170	190	205	215	220

$$Md. = 2,5 + \frac{1}{30}(110 - 90) = 3,17$$

esos son los valores escalares asignados a las campañas de los distintos productos. La peor campaña es la correspondiente al producto número 3 y, en relación a las otras 2 hemos de decir que ninguna de ellas es suficientemente buena. Sus valores escalares están en torno a los valores medios.

3. Hay que calcular la puntuación total de cada sujeto en los 5 ítems.

3.1. El sujeto que manifestó una actitud más desfavorable fue el sujeto 4 (tuvo una puntuación de 12 puntos), el que manifestó una actitud más favorable fue el sujeto 9 que obtuvo 22 puntos. Dado que la puntuación máxima por cada ítem es 5 y hay 5 ítems en la escala, la puntuación máxima de la escala sería 25 puntos y la mínima 5.

3.2. No es demasiado homogéneo ya que las puntuaciones oscilan desde 12 a 22. Lo que sí es cierto es que ninguno de los sujetos muestra una actitud muy desfavorable ya que se sitúan desde las puntuaciones medias hacia las puntuaciones altas

4. A partir de las medias escalares podemos averiguar las puntuaciones factoriales para cada concepto y grupo, teniendo en cuenta las escalas que saturan en cada una de las dimensiones.

Así, por ejemplo, la puntuación factorial del concepto GUERRA, en la dimensión evaluativa, y para el grupo A, se obtendrá sumando algebraicamente las medias correspondientes a las escalas bipolares que definen esta dimensión y dividiendo esta suma por el número de escalas. Dado que tenemos tres escalas bipolares que saturan en esta dimensión, cuyas medias son (-3), (-2) y (-3), la puntuación factorial buscada será:

$$PF = \frac{-3 - 2 - 3}{3} = -2,67$$

La puntuación factorial del concepto GUERRA, en la dimensión evaluativa, pero en el grupo B será:

$$PF = \frac{-2 - 1 - 2}{3} = -1,67$$

Del mismo modo iríamos calculando las puntuaciones factoriales de cada concepto en cada una de las dimensiones. Los resultados obtenidos, aparecen recogidos en el siguiente cuadro:

	EVALUATIVA		**ACTIVIDAD**		**POTENCIA**	
	G.A	G.B	G.A	G.B	G.A	G.B
GUERRA	-2,67	-1,67	0,67	2,67	2,67	2,67
AMOR	3	3	2	1,67	2,33	1,67

A pesar de que la información que ofrecen estas puntuaciones no es demasiado eficiente desde el punto de vista conceptual, podemos tener una idea acerca de los significados culturales de los conceptos. Así, por ejemplo, podríamos decir que en el grupo A la GUERRA es valorada de forma bastante negativa (-2,67), casi indiferente en cuanto a la dimensión de actividad (0,67) y muy potente (2,67); mientras que el grupo B, la valora de forma menos negativa (-1,67), bastante potente y también bastante activa.

5. Soluciones a los ejercicios conceptuales

1. La afirmación es correcta.

   Cada vez que se presenta un estímulo a un sujeto se produce un proceso discriminante mediante el cual asigna un valor al estímulo.

2. La afirmación es incorrecta.

   Debido a una serie de factores: motivacionales, ambientales, personales, etc. cuando a un sujeto se le presenta un mismo estímulo, puede suscitar en él procesos discriminantes distintos y, por lo tanto, los valores asignados por el sujeto al estímulo, a través de los distintos procesos discriminantes, pueden variar.

3. La afirmación es incorrecta.

   La desviación típica es un índice del grado de ambigüedad que suscita el estímulo en el sujeto (o sujetos); a medida que el grado de ambigüedad es mayor, la desviación típica será mayor también.

4. La afirmación es correcta

5. La afirmación es correcta.

   Ese es uno de los supuestos del modelo de Thurstone

6. La afirmación es correcta.

   Se trata de un método de escalamiento de estímulos. Ahora bien, una vez construida la escala y asignados los valores escalares a los estímulos (ítems) en función del grado de actitud que lleven implícito, se puede aplicar la escala a los sujetos para escalarlos en función de la actitud que manifiesten.

7. La afirmación es incorrecta.

   Los sujetos no pueden asignar directamente el valor escalar a los distintos estímulos ya que los valores escalares son las medias de las distribuciones discriminativas.

8. La afirmación es correcta.

9. La afirmación es correcta.

10. La afirmación es correcta.

11. La afirmación es incorrecta.

    Las escalas de Likert se desarrollaron para la medida de las actitudes y para poder diferenciar a los sujetos respecto a ellas.

12. La afirmación es incorrecta.

    Las escalas de likert no utilizan la *prueba de jueces* para la asignación de valores escalares a los ítems. Es el propio investigador el que, siguiendo la dirección de la variable de actitud medida, asigna directamente los valores numéricos a las distintas categorías de respuesta de los ítems.

13. La afirmación es correcta.

    En efecto, se trata de una forma de abordar el problema de la medida de las actitudes basándose en el principio de que un mismo objeto o estímulo tiene distinto significado para las personas que lo evalúan y, por lo tanto, la actitud que cada una de ellas muestre hacia dicho estímulo dependerá del significado que tenga para ella.

14. La afirmación es correcta.

    En todo Diferencial Semántico hay dos elementos fundamentales: los conceptos y las escalas bipolares. Los conceptos representan los objetos o estímulos a evaluar y las escalas bipolares el «instrumento» que se utiliza para evaluarlos.

15. La afirmación es incorrecta.

    El significado semántico de los conceptos puede ser explicado por más de tres dimensiones. Estas, vendrán determinadas por las escalas bipolares que se utilicen en su evaluación, y para averiguar cuales son las dimensiones subyacentes será necesario utilizar alguna de las técnicas que hemos explicado anteriormente.

16. La afirmación es incorrecta.

    No basta con una escala bipolar por cada una de las dimensiones para evaluar de forma adecuada el significado semántico de los conceptos. En principio podría valer la regla de que «cuantas más escalas mejor»; no obstante, y como norma general, se suelen utilizar cuatro a seis escalas bipolares por cada una de las dimensiones.

17. La afirmación es correcta.

    A lo largo del continuo psicológico se encuentran entrelazados los sujetos y los ítems.

18. La afirmación es correcta.

    Si el ajuste fuera perfecto la matriz resultante sería una matriz triangular.

19. La afirmación es incorrecta.

    El coeficiente de reproductividad no puede ser negativo, su valor oscila entre 0 y 1. Será cero cuando el número de errores coincida con el número de juicios emitidos y será uno cuando no haya ningún error.

20. La afirmación es correcta.

12. BIBLIOGRAFÍA COMPLEMENTARIA

A lo largo de estas obras se pueden encontrar, de forma extensa, todas las técnicas que se han incluido en este tema.

Barbero, M.I. (2007). *Métodos de elaboración de escalas.* Madrid: UNED.

Summers, G. F. (1976). *Medición de actitudes.* Madrid: Trillas

Wainerman, C. et al. (1976). *Escalas de Medición en las Ciencias Sociales.* Madrid: Nueva Visión

Yela, M. (1966). El método de las comparaciones binarias y la construcción de escalas psicológicas. *Revista de Psicología General y Aplicada,* 21, 89, 659-690.

Este artículo es uno de los más claros sobre la construcción de escalas mediante la Ley del Juicio Comparativo.

Parte II

EVALUACIÓN DE LAS PROPIEDADES MÉTRICAS DE LOS INSTRUMENTOS DE MEDICIÓN PSICOLÓGICA

TEMA 4

LA FIABILIDAD DE LAS PUNTUACIONES

Enrique Vila Abad

SUMARIO

1. ORIENTACIONES DIDÁCTICAS

En los temas anteriores se han expuesto los principios básicos para la construcción de tests y las principales técnicas de construcción de escalas de actitudes; se cumple así la primera etapa de la construcción de los instrumentos de medición psicológica. Hasta el momento, se cuenta con una prueba piloto que se ha aplicado a una muestra de sujetos a los que se les han asignado sus puntuaciones correspondientes. Vamos a comenzar ahora el estudio de la segunda parte del proceso, la evaluación de la calidad métrica de la prueba piloto y la construcción del instrumento de medición definitivo.

Hemos intentado aclarar, en cierta medida, los distintos términos utilizados en relación con estos instrumentos: tests, escalas, cuestionarios, etc.; sin embargo, a partir de este momento, y teniendo en cuenta que la forma de llevar a cabo la evaluación de la calidad métrica es la misma, vamos a seguir las mismas normas que en los *Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014)* que utilizan el término *test* para referirse a todos estos instrumentos de evaluación.

Esta fase de evaluación de la calidad del test debería comenzar por el análisis de la calidad de los ítems ya que, como se ha comentado con anterioridad, dado que los ítems son las unidades elementales del test, difícilmente se podrá contar con un buen test si los ítems que lo forman son de mala calidad. Durante el proceso de construcción de la prueba inicial se ha explicado la forma de llevar a cabo una parte del análisis de los ítems a partir de la revisión crítica del contenido de los mismos por un grupo de expertos o jueces; sin embargo, quedaría por hacer otro tipo de análisis que no estuviera basado en juicios subjetivos, sino un análisis objetivo basado en las respuestas que han emitido los sujetos a los ítems. Dado que para llevar a cabo este tipo de análisis es necesario que nuestros alumnos se hayan familiarizado con los conceptos de validez y fiabilidad, entre otros, el tema dedicado al análisis de los ítems se expondrá más adelante.

Una vez evaluada la calidad de los ítems del test y eliminados aquellos que no se consideran adecuados, el paso siguiente será la evaluación de la calidad global del test que incluye, entre

otras cosas, la evaluación de la precisión y estabilidad de las medidas (fiabilidad) y la pertinencia de las inferencias realizadas a partir de las puntuaciones obtenidas (validez).

En este tema se analiza el problema de la fiabilidad y precisión de la medida, tratando de encontrar respuesta a la pregunta de hasta qué punto las puntuaciones obtenidas por los sujetos en la prueba que se les ha aplicado están afectadas por errores de medida y en qué cuantía. El tema siguiente está dedicado al estudio de la fiabilidad desde la perspectiva de los tests referidos al criterio (TRC).

Comenzamos el tema con una alusión al problema del error de medida, centrándonos en los errores aleatorios y en cómo el modelo lineal propuesto por Spearman intenta buscar soluciones a esta cuestión. Seguidamente presentamos los distintos tipos de errores aleatorios con los que nos podemos encontrar al aplicar un instrumento de medición.

A continuación, introducimos la definición, dentro de la Teoría Clásica de los Tests, del coeficiente de fiabilidad, haciendo mención de los distintos factores que pueden influir en su cuantía como pueden ser la longitud del test y las características de la muestra a la que se aplica, y explicando la necesidad de establecer procedimientos empíricos que nos permitan estimarlo: *el método de las formas paralelas, el método test-retest y los métodos basados en la consistencia interna del test;* indicando cómo se deben interpretar los coeficientes obtenidos. A continuación se presentan tres procedimientos que permitirán estimar el nivel real del sujeto en el rasgo o característica que mide el test (su puntuación verdadera).

Al estudiar el tema se recomienda que los alumnos hagan hincapié en los siguientes puntos básicos:

- Conocer los supuestos básicos del modelo lineal de Spearman así como las deducciones que se puedan hacer a partir de esos supuestos.
- Tener muy claros los conceptos de error de medida y fiabilidad.
- Saber diferenciar los distintos tipos de error de medida.
- Conocer la influencia que pueden tener en el coeficiente de fiabilidad factores como la longitud del test y la variabilidad de la muestra de sujetos a los que se aplica.
- Conocer los procedimientos empíricos para estimar el coeficiente de fiabilidad.
- Diferenciar entre la fiabilidad como estabilidad temporal de las puntuaciones obtenidas en el test y como consistencia interna de los ítems del test.
- Diferenciar entre los distintos procedimientos para estimar la puntuación verdadera de un sujeto en un test.

Nota: Para aquellos lectores interesados, al final del tema se incluye un Apéndice en el que se ofrecen las demostraciones de algunas de las fórmulas más significativas que irán apareciendo a lo largo del mismo.

2. EL PROBLEMA DEL ERROR DE MEDIDA

Uno de los requisitos fundamentales en cualquier teoría de la medición es la fiabilidad y precisión de los instrumentos utilizados para medir una determinada característica. La medición en Psicología no está exenta de este requisito y debemos contar con instrumentos que sean fiables y, por consiguiente, libres en la medida de lo posible, de errores de medida. El concepto de *error de medida* es un concepto básico en Psicometría.

> Se define el error de medida como la diferencia entre la puntuación empírica obtenida por un sujeto en un test y su puntuación verdadera, entendiendo por test cualquier instrumento de medición psicológica.

Si aplicáramos «*n*» veces un test a un mismo sujeto, con la finalidad de determinar su capacidad en una determinada característica, es casi seguro que las puntuaciones obtenidas por ese sujeto serían muy parecidas pero nunca iguales observándose que, en algunos casos, el valor de la puntuación empírica estará por encima de la puntuación verdadera del sujeto, la que realmente indica la capacidad que tiene, y en otros por debajo. En cualquier caso será responsabilidad del investigador construir pruebas que den lugar al mínimo error de medida posible, y que la puntuación obtenida proporcione el mayor grado de información real sobre la característica objeto de estudio.

A veces, los errores de medida no son debidos al propio instrumento de medición sino a cambios que operan en el propio sujeto y que pueden ser atribuidos a diversas razones: su motivación cuando realiza la prueba, que conteste al azar algunos de los ítems, las condiciones físicas en que se encuentre, etc. Éstos son errores de *carácter aleatorio* e impredecibles, con los que hay que contar y tratar de controlar para que no interfieran de manera significativa en las predicciones que podamos hacer acerca de su capacidad. Son los errores de los que se va a ocupar la fiabilidad. En el apartado 6 veremos con más detalle los distintos tipos de errores de medida que existen.

De lo dicho hasta ahora se puede deducir, en primer lugar, que si aplicamos repetidas veces un mismo test a un sujeto, lo más probable es que obtengamos puntuaciones distintas en las diferentes aplicaciones y, en segundo lugar, que cuando un sujeto obtiene una puntuación en un test, dicha puntuación estará afectada por errores de medida. Este hecho nos lleva a plantearnos la siguiente pregunta: ¿cómo podemos saber cuál es el valor real del sujeto en la característica que estamos estudiando? Para ello, hay que acudir a alguna de las teorías que se han ido desarrollando y que nos van a proporcionar los medios para hacer estimaciones acerca de la cuantía de error que afecta a las puntuaciones empíricas y acerca del verdadero nivel del sujeto (o sujetos) en la característica que se está midiendo.

Dado que este texto está dedicado, fundamentalmente, a la Teoría Clásica de los Tests, el modelo (teoría) que se estudiará es el modelo lineal propuesto por Spearman. Este modelo establece que la puntuación empírica obtenida por los sujetos cuando se les aplica un test es función lineal de su puntuación verdadera en el rasgo que se intenta medir y un componente de error, tal y como se especificará en el siguiente apartado.

3. El MODELO LINEAL DE SPEARMAN

El modelo lineal de Spearman, establece que la puntuación empírica obtenida por un sujeto en un test (*X*) puede considerarse como una combinación lineal de dos componentes: por una parte, la puntuación verdadera (*V*) de ese sujeto en el rasgo que mide el test, y por otra, el error de medida (*E*) que la afecta. Así pues, podemos establecer la ecuación del modelo en los siguientes términos:

$$X = V + E \qquad [4.1]$$

Como se puede deducir de esta expresión, si aplicamos un test a un sujeto la puntuación que obtenga en el test, no coincidirá con el valor de la puntuación verdadera. Como en cualquier proceso de medición hemos de tener en cuenta la presencia del error de medida cometido.

El modelo asume una serie de supuestos:

Primer supuesto. La puntuación verdadera (*V*) es la esperanza matemática de la puntuación empírica (*X*). Esto quiere decir que si a un sujeto se le pasara un número infinito de veces un mismo test, y suponiendo que las aplicaciones fueran independientes entre sí de manera que la puntuación obtenida por dicho sujeto en una de las aplicaciones no estuviera influyendo en la obtenida en las demás, la media de todas las puntuaciones observadas (*X*) sería la puntuación verdadera del sujeto.

$$V = E(X) \qquad [4.2]$$

Segundo supuesto. La correlación existente entre las puntuaciones verdaderas de «*n*» sujetos en un test y los errores de medida es igual a cero. Es decir, no existe relación entre los errores de medida y las puntuaciones verdaderas.

$$r_{ve} = 0 \qquad [4.3]$$

Tercer supuesto. La correlación entre los errores de medida ($r_{e_1e_2}$) que afectan a las puntuaciones de los sujetos en dos tests diferentes (X_1 y X_2) es igual a cero. Si «e_1» representa los errores de medida de las puntuaciones de «*n*» sujetos en el test 1 y «e_2» representa los errores de medida de las puntuaciones de los mismos sujetos en el test 2 el supuesto implica que no existe ninguna razón para presuponer que los errores de medida cometidos en un test vayan a influir, positiva o negativamente, en el otro test, siempre y cuando los tests se apliquen correctamente.

$$r_{e_1e_2} = 0 \qquad [4.4]$$

A partir de estos tres supuestos del modelo se pueden hacer las siguientes deducciones:

a) El error de medida se define como la diferencia entre la puntuación empírica obtenida por un sujeto y su puntuación verdadera.

$$E = X - V \qquad [4.5]$$

b) La esperanza matemática de los errores de medida es cero.

$$E(e) = 0 \qquad [4.6]$$

c) La media de las puntuaciones empíricas es igual a la media de las puntuaciones verdaderas.

$$\bar{X} = \bar{V} \qquad [4.7]$$

d) La covarianza entre las puntuaciones verdaderas y los errores es igual a cero.

$$Cov\ (V, E) = 0 \qquad [4.8]$$

e) La varianza de las puntuaciones empíricas es igual a la suma de la varianza de las puntuaciones verdaderas más la varianza de los errores.

$$S_x^2 = S_v^2 + S_e^2 \qquad [4.9]$$

f) La covarianza entre las puntuaciones empíricas y las verdaderas es igual a la varianza de las puntuaciones verdaderas.

$$Cov(X,V) = S_v^2 \qquad [4.10]$$

g) La correlación entre las puntuaciones empíricas y los errores es igual al cociente entre la desviación típica de los errores y la desviación típica de las puntuaciones empíricas.

$$r_{xe} = \frac{S_e}{S_x} \qquad [4.11]$$

h) La covarianza entre las puntuaciones empíricas de dos tests es igual a la covarianza entre las puntuaciones verdaderas.

$$Cov\ (X_1, X_2) = Cov\ (V_1, V_2) \qquad [4.12]$$

De donde se deduce que la covarianza entre las puntuaciones empíricas obtenidas por una muestra de sujetos en las dos aplicaciones de un test es igual a la varianza de las puntuaciones verdaderas, dado que al ser el mismo test el que se aplica en dos ocasiones distintas la puntuación verdadera es la misma y la covarianza de una variable consigo misma es igual a la varianza. (Véase el punto 4.12 del Apéndice al final del tema).

4. TESTS PARALELOS. CONDICIONES DE PARALELISMO

Si a una misma muestra de sujetos se le aplican dos tests, X y X', podemos considerar que son paralelos si, además de cumplirse los supuestos anteriores, se cumplen las dos condiciones siguientes:

1. Las puntuaciones verdaderas de los sujetos son iguales en ambos tests.

   Según el modelo lineal podemos establecer:

$$X = V + E$$
$$X' = V + E'$$

2. La varianza de los errores de medida es la misma en ambos tests:

$$S_e^2 = S_{e'}^2$$

De las condiciones de paralelismo enunciadas podemos sacar una serie de deducciones importantes dentro del modelo clásico.

a) La media de las puntuaciones empíricas obtenidas en dos tests supuestamente paralelos es la misma.

Teniendo en cuenta que la esperanza matemática de los errores de medida es cero y que las puntuaciones verdaderas de los sujetos son iguales en ambos tests, podemos concluir la existencia de igualdad entre las medias de las puntuaciones empíricas.

$$\bar{X} = \bar{X}'$$
$$\bar{X} = \bar{V} + \bar{E} = \bar{V}$$
$$\bar{X}' = \bar{V} + \bar{E}' = \bar{V}$$

b) Las varianzas de las puntuaciones empíricas obtenidas en dos tests paralelos son iguales.

$$S_X^2 = S_{X'}^2$$
$$S_X^2 = S_V^2 + S_e^2$$
$$S_{X'}^2 = S_V^2 + S_{e'}^2$$

Teniendo en cuenta, por definición de tests paralelos, que la varianza de los errores es la misma, podemos concluir que las varianzas de las puntuaciones empíricas son iguales.

c) La correlación entre las puntuaciones empíricas obtenidas en dos tests paralelos ($r_{XX'}$) es igual al cuadrado de la correlación entre las puntuaciones empíricas y las puntuaciones verdaderas (r_{XV}^2) o bien, al cociente entre la varianza de las puntuaciones verdaderas y la varianza de las puntuaciones empíricas.

$$r_{XX'} = r_{XV}^2 = \frac{S_V^2}{S_X^2} \qquad [4.13]$$

d) Dados dos o más tests paralelos, las intercorrelaciones entre cada dos de ellos son iguales.

$$r_{X_1X_2} = r_{X_1X_3} = r_{X_2X_3} = \ldots\ldots = r_{X_jX_k} \qquad [4.14]$$

5. INTERPRETACIÓN TEÓRICA DEL COEFICIENTE DE FIABILIDAD

Definimos el coeficiente de fiabilidad de un test, como:

> ... la correlación entre las puntuaciones empíricas obtenidas por una muestra de sujetos en dos formas paralelas del test.

Se puede expresar también como el cociente entre la varianza de las puntuaciones verdaderas y la varianza de las puntuaciones empíricas.

$$r_{xx'} = \frac{S_V^2}{S_X^2} \qquad [4.15]$$

y se pude interpretar, por lo tanto, como la proporción de la varianza de las puntuaciones empíricas de los sujetos que se debe a la varianza de las puntuaciones verdaderas, o lo que es lo mismo, la proporción de varianza verdadera que hay en la varianza empírica. A medida que dicha proporción aumenta, disminuye el error de medida. Si $r_{xx'} = 1$, el error de medida es cero lo que implica una fiabilidad perfecta del test. Sin embargo, a medida que dicha proporción disminuye se produce un incremento en el error de medida. En el caso de que $r_{xx'} = 0$, la varianza de los errores de medida sería igual a la varianza de las puntuaciones empíricas.

EJEMPLO:

Calcular el coeficiente de fiabilidad de un test de razonamiento abstracto, sabiendo que la varianza verdadera de dicho test es el 80% de su varianza empírica.

$$r_{xx'} = \frac{S_V^2}{S_X^2} = \frac{0,80S_X^2}{S_X^2} = 0,80$$

es decir el 80% de la varianza de las puntuaciones empíricas es verdadera medida del rasgo.

A partir de la expresión (4.13) se puede inferir que:

$$\text{Si } r_{xx'} = r_{xv}^2 \Rightarrow r_{xv} = \sqrt{r_{xx'}} \qquad [4.16]$$

Al término r_{xv} se le denomina *índice de fiabilidad de un test.*

El coeficiente de fiabilidad de un test se puede expresar también en función de la varianza de los errores:

$$r_{xx'} = 1 - \frac{S_e^2}{S_x^2} = 1 - r_{xe}^2 \qquad [4.17]$$

Así mismo, es fácilmente deducible que:

$$r_{xe} = \frac{S_e}{S_x} \Rightarrow r_{xe} = \sqrt{1 - r_{xx'}} \qquad [4.18]$$

Es decir, la correlación entre las puntuaciones empíricas y los errores de medida se puede obtener a partir de la correlación entre las puntuaciones obtenidas por los sujetos en las dos formas paralelas de un test. El término $\frac{S_e}{S_x}$ representa la proporción de la desviación típica de las puntuaciones empíricas de los sujetos en el test que se debe a la desviación típica de los errores y, como vemos, esa proporción se puede estimar a partir del coeficiente de fiabilidad del test.

Resumiendo, podemos decir que el coeficiente de fiabilidad definido según el modelo clásico de Spearman como la correlación entre las puntuaciones obtenidas por una muestra de sujetos en dos tests paralelos, nos proporciona información para poder estimar la cuantía del error de medida.

6. TIPOS DE ERRORES DE MEDIDA

En este apartado haremos alusión a diferentes tipos de errores: el de medida, el de estimación, el de sustitución y el de predicción.

— *Error de medida.*

Como ya se ha dicho, el error de medida es la diferencia entre la puntuación empírica de un sujeto y su puntuación verdadera.

$$E = X - V$$

A la desviación típica de los errores de medida se le denomina *error típico de medida* y se expresa como:

$$S_e = S_x\sqrt{1-r_{xx'}} \qquad [4.19]$$

Cuando se calcula el error de medida obtenemos una medida individual del error que se comete; es decir, una medida individual de la precisión del test. Esta medida nos indica la diferencia que existe entre la puntuación que un sujeto ha obtenido en un test y el nivel real de dicho sujeto en la variable que medimos con dicho test; es decir, su puntuación verdadera. Cuando calculamos el error típico de medida estamos llevando a cabo una medida grupal del error puesto que se calcula para todos los sujetos de la muestra.

Este mismo razonamiento es válido para los distintos tipos de error que se exponen a continuación.

— *Error de estimación de la puntuación verdadera.*

Se denomina error de estimación de la puntuación verdadera a la diferencia entre la puntuación verdadera de un sujeto y la puntuación verdadera pronosticada mediante el modelo de regresión.

$$E = V - V'$$

Definimos el *error típico de estimación de la puntuación verdadera*, como la desviación típica de los errores de estimación y viene expresado como:

$$S_{vx} = S_x\sqrt{1-r_{xx'}}\sqrt{r_{xx'}} = S_e\sqrt{r_{xx'}} \qquad [4.20]$$

— *Error de sustitución*

Se define el error de sustitución como la diferencia entre las puntuaciones obtenidas por un sujeto en un test y las obtenidas en otro test paralelo. Es decir, el error que se cometería al sustituir las puntuaciones obtenidas en un test X_1 por las obtenidas en un test paralelo X_2.

$$e = X_1 - X_2$$

Definimos el *error típico de sustitución*, como la desviación típica de los errores de sustitución y viene expresado como:

$$S_{X_1-X_2} = S_x\sqrt{1-r_{xx'}}\sqrt{2} \qquad [4.21]$$

— *Error de predicción.*

Se define el error de predicción como la diferencia entre las puntuaciones obtenidas por un sujeto en un test (X_1) y las puntuaciones pronosticadas en ese mismo test (X_1') a partir de una forma paralela X_2.

$$e = X_1 - X_1'$$

La puntuación X_1' se obtiene mediante la recta de regresión de X_1 sobre X_2:

$$X'_1 = r_{12}\frac{S_{x_1}}{S_{x_2}}\left(X_2 - \bar{X}_2\right) + \bar{X}_1 \qquad [4.22]$$

Definimos el *error típico de predicción*, como la desviación típica de los errores de predicción y viene expresado como:

$$S_{e_p} = S_x\sqrt{1-r_{xx'}}\sqrt{1+r_{xx'}} = S_e\sqrt{1+r_{xx'}} \qquad [4.23]$$

7. FACTORES QUE AFECTAN A LA FIABILIDAD

La fiabilidad de un test depende de factores como la variabilidad del grupo al que se aplica, la longitud del propio test, las características de los ítems que lo configuran, etc. En este apartado estudiaremos los dos primeros aspectos y el tercero será abordado más adelante en otro tema dedicado específicamente al estudio de la calidad métrica de los ítems.

7.1. Longitud del test

Uno de los factores que influyen en la fiabilidad de un test es su longitud, es decir, el número de ítems que lo componen. Cuantos más ítems representativos del rasgo a medir se utilicen mayor será la información que obtengamos acerca del atributo que estemos estudiando. Cabe pensar que también será menor el error que cometamos al tratar de estimar la puntuación verdadera de un sujeto y, por lo tanto, la fiabilidad del test tenderá a incrementarse. Una forma de poder aumentar la fiabilidad del test es aumentar su longitud. A veces, si un test es demasiado largo puede ser interesante averiguar cuál sería su fiabilidad si se le disminuyera el número de ítems. Si esta disminución de la fiabilidad no es muy elevada puede ser más aconsejable utilizar el test más corto.

La relación existente entre la fiabilidad de un test y su longitud, siempre y cuando los ítems a añadir sean paralelos a los que ya tenía el test original, se puede evaluar mediante la ecuación de Spearman-Brown.

$$R_{xx} = \frac{nr_{xx}}{1 + nr_{xx} - r_{xx}} = \frac{nr_{xx}}{1 + (n-1)r_{xx}} \qquad [4.24]$$

donde:

R_{xx} = coeficiente de fiabilidad del test alargado o acortado.

r_{xx} = coeficiente de fiabilidad del test inicial.

n = número de veces que se ha alargado o acortado el test.

$n = \frac{EF}{EI}$, siendo EF el número de elementos finales del test y EI el número de elementos iniciales del test.

Esta expresión (4.24), es la que definimos como ecuación general de Spearman-Brown y hace referencia al caso en que se quiera aumentar o disminuir la longitud del test inicial «n» veces.

Todo lo que acabamos de decir, es igual de válido en el caso de reducir la longitud del test, con la salvedad de que «n» será siempre menor que 1.

Nota: Téngase en cuenta que «n» no es el número de ítems que se añaden o se eliminan del test original, sino que hace referencia al número de veces que se aumenta o disminuye la longitud del test.

EJEMPLO:

Supongamos, que se aplica un test de percepción visual compuesto por 50 ítems a una muestra de sujetos y se obtiene un coeficiente de fiabilidad de 0,60. Veamos lo que sucede al incrementar n veces la longitud del test:

Para $n = 2$; $R_{xx} = \frac{nr_{xx}}{1 + (n-1)r_{xx}} = \frac{2 \cdot 0,60}{1 + 0,60} = 0,75$ para $n = 3$; $R_{xx} = \frac{1,80}{1 + 1,20} = 0,82$

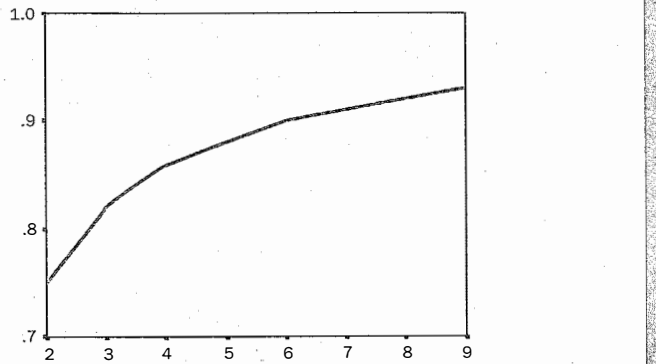
Para $n = 4$; $R_{xx} = \frac{2,40}{1 + 1,80} = 0,86$ para $n = 5$; $R_{xx} = \frac{3}{1 + 2,40} = 0,88$

Para $n = 6$; $R_{xx} = \frac{3,60}{1 + 3} = 0,90$ para $n = 7$; $R_{xx} = \frac{4,20}{1 + 3,6} = 0,91$

Para $n = 8$; $R_{xx} = \frac{4,80}{1 + 4,20} = 0,92$ para $n = 9$; $R_{xx} = \frac{5,40}{1 + 4,80} = 0,93$

Como se puede apreciar en el gráfico 4.1, a medida que aumenta el número de ítems paralelos aumenta el coeficiente de fiabilidad del test, aunque no de una manera proporcional. Se puede observar que a partir de un determinado valor de n no se producen incrementos significativos en la fiabilidad del test. Como consecuencia de esto nos podemos preguntar: ¿cuánto habría que alargar o acortar un test para obtener un determinado coeficiente de fiabilidad? y, en segundo lugar, ¿hasta qué punto es razonable dicho aumento?

GRÁFICO 4.1
Relación entre la fiabilidad y la longitud del test



De nuevo encontramos la respuesta a esta pregunta a través de la ecuación de Spearman-Brown, ya que despejando «n» tendremos:

$$n = \frac{R_{xx} - r_{xx}R_{xx}}{r_{xx} - r_{xx}R_{xx}} = \frac{R_{xx}(1-r_{xx})}{r_{xx}(1-R_{xx})}$$

Supongamos que con los datos del ejemplo anterior queremos aumentar la fiabilidad del test hasta obtener un valor de 0,93. Aplicando la expresión anterior tenemos:

$$n = \frac{R_{xx}(1-r_{xx})}{r_{xx}(1-R_{xx})} = \frac{0,93(1-0,60)}{0,60(1-0,93)} = 8,85 \cong 9$$

Para conseguir ese coeficiente de fiabilidad sería necesario hacer, aproximadamente, 9 veces más largo el test original de 50 ítems. Es decir, el nuevo test tendría una longitud de:

$$n = \frac{EF}{EI}; \quad EF = n \cdot EI = 9 \cdot 50 = 450 \text{ ítems}$$

lo que implicaría añadir 400 ítems al test original. Obviamente, añadir al test 400 ítems no parece una solución razonable al problema y debemos plantearnos otras alternativas como, por ejemplo, revisar el objetivo para el que se construye el test, mejorar los ítems y analizar cuál es el valor de precisión aceptable para dicho objetivo (si se hubiera tomado el valor exacto de 8,85 habría que haber añadido 392,5 ítems es decir 393, lo cual tampoco sería una solución razonable).

En ocasiones puede que estemos interesados en saber si es posible reducir el número de ítems de un test y que el nuevo coeficiente de fiabilidad sea lo suficientemente aceptable como para no perder demasiada información respecto al atributo objeto de estudio. Esta situación se da cuando consideramos el número de ítems excesivo. Supongamos, por ejemplo un test compuesto de 100 ítems y un coeficiente de fiabilidad de 0,85 (r_{xx}). Supongamos que para nuestros objetivos un coeficiente de fiabilidad de 0,75 (R_{xx}) es admisible. La pregunta que nos formularíamos sería cuántos elementos debemos eliminar del test original para obtener ese coeficiente de fiabilidad. En este caso:

$$n = \frac{R_{xx}(1-r_{xx})}{r_{xx}(1-R_{xx})} = \frac{0,75(1-0,85)}{0,85(1-0,75)} = \frac{0,11}{0,21} = 0,52$$

$$EF = n \cdot EI = 0,52 \times 100 = 52$$

luego tendríamos que eliminar 100 – 52 = 48 ítems.

7.2. Variabilidad de la muestra

La fiabilidad de un test también depende de las características de la muestra a la que se aplica. Un test puede presentar tantos coeficientes de fiabilidad como muestras distintas en las que se calcule. El coeficiente de fiabilidad puede variar en función de la mayor o menor homogeneidad del grupo, siendo menor cuanto más homogéneo sea; es decir,cuanto más pequeña sea la desviación típica de las puntuaciones empíricas obtenidas por los sujetos en el test. Recordemos que hemos definido el coeficiente de fiabilidad como la correlación entre dos formas paralelas de un test y la correlación viene afectada por la variabilidad del grupo. Por lo tanto, es importante saber hasta qué punto la fiabilidad de un test se ve afectada por dicha variabilidad.

Supongamos dos grupos de sujetos 1 y 2. Partiendo del supuesto de que el error típico de medida de un test se mantiene constante, independientemente de la variabilidad del grupo en que se aplique, podemos establecer la siguiente igualdad:

$$S_{e_1}^2 = S_{e_2}^2$$

Por tanto, teniendo en cuenta que, $S_e^2 = S_x^2(1-r_{xx})$ podemos establecer la igualdad: $S_1^2(1-r_{11}) = S_2^2(1-r_{22})$ y despejando:

$$\boxed{r_{22} = 1 - \frac{S_1^2}{S_2^2}(1-r_{11})} \qquad [4.25]$$

donde:

S_1^2 = varianza empírica de las puntuaciones en el grupo 1.

S_2^2 = varianza empírica de las puntuaciones en el grupo 2.

r_{11} = coeficiente de fiabilidad en el grupo 1.

r_{22} = coeficiente de fiabilidad en el grupo 2.

EJEMPLO:

Se ha aplicado un test a una muestra de sujetos en la que la desviación típica de las puntuaciones empíricas obtenidas es igual a 20 y la razón entre la desviación típica de los errores y la desviación típica de las puntuaciones empíricas es 0,40. Aplicado el test a otra muestra de sujetos en la que la desviación típica de las puntuaciones empíricas es igual a 10, ¿cuál sería el valor del coeficiente de fiabilidad del test?

Datos: $S_{x1} = 20$; $S_{x2} = 10$; $\frac{S_{e1}}{S_{x1}} = 0,40$

$$r_{11} = 1 - \frac{S_e^2}{S_x^2} = 1 - 0,16 = 0,84$$

$$S_1^2(1 - r_{11}) = S_2^2(1 - r_{22});\ 400(1 - 0,84) = 100(1 - r_{22});\ 64 = 100 - 100\ r_{22};\ r_{22} = 0,36$$

Como se puede apreciar, al reducir la variabilidad de las puntuaciones empíricas en el segundo grupo, se reduce el coeficiente de fiabilidad. Asimismo se puede observar que el valor del error típico de medida permanece constante (lo cuál es lógico puesto que se ha partido de ese supuesto).

$$S_{e1} = S_{x1}\sqrt{1 - r_{xx}} = 20\sqrt{1 - 0,84} = 8$$

$$S_{e2} = S_{x2}\sqrt{1 - r_{xx}} = 10\sqrt{1 - 0,36} = 8$$

8. LA FIABILIDAD COMO EQUIVALENCIA Y COMO ESTABILIDAD DE LAS MEDIDAS

Un test debe cumplir dos requisitos básicos. En primer lugar debe medir el rasgo que realmente pretende medir (es decir, ser válido) y, en segundo lugar, las puntuaciones empíricas obtenidas al aplicar el test deben ser estables y precisas. La precisión hace referencia, como ya hemos apuntado anteriormente, a la necesidad de que, en la medida de lo posible, las puntuaciones obtenidas estén libres de errores. La estabilidad se refiere a que cuando se evalúa un rasgo con el mismo test en distintas ocasiones y bajo condiciones lo más parecidas posibles, siempre y cuando el rasgo estudiado no haya cambiado, se deberán obtener unos resultados similares. Este segundo requisito, referido a la reproductividad de unos resultados en condiciones similares, es lo que definimos como la fiabilidad del test, entendida como estabilidad de las medidas. En definitiva, lo que pretendemos es poder establecer el grado de acuerdo entre las puntuaciones obtenidas por los sujetos en distintas aplicaciones.

En este apartado, nos centraremos en dos métodos basados en la estabilidad de las medidas para el cálculo del coeficiente de fiabilidad, métodos que constituyen una aplicación directa de la definición de correlación entre formas paralelas:

— Método de las formas paralelas

— Método test-retest

Existen otras formas de abordar el cálculo de la fiabilidad de un test, como veremos en el siguiente apartado, basadas en la consistencia interna del test.

8.1. Método de las formas paralelas

La forma de proceder, según este método, sería: primero, construir dos formas paralelas de un test X y X', en segundo lugar, aplicar las dos formas del test a una muestra de sujetos lo suficientemente amplia como para que sea representativa de la población a la que va dirigido el test y, en tercer lugar, calcular el coeficiente de correlación de Pearson entre las puntuaciones de los sujetos en ambas formas.

$$r_{XX'} = r_{x_1x_2} = \frac{N\sum X_1X_2 - \sum X_1 \sum X_2}{\sqrt{\left[N\sum X_1^2 - (\sum X_1)^2\right]\left[N\sum X_2^2 - (\sum X_2)^2\right]}} \qquad [4.26]$$

donde: X_1 y X_2 corresponden a las puntuaciones obtenidas por los sujetos en cada una de las formas aplicadas.

El coeficiente de fiabilidad así obtenido recibe también el nombre de *coeficiente de equivalencia*, haciendo referencia al grado en que ambas formas son equivalentes.

El método de las formas paralelas presenta la ventaja de que, si ambas formas son aplicadas en el mismo momento se tiene un mayor control de las condiciones en que los sujetos realizan las pruebas. Este método presenta el inconveniente de la dificultad que supone la construcción de dos formas que sean paralelas.

8.2. Método test-retest

Con este método se aplica el mismo test en dos ocasiones diferentes a una misma muestra de sujetos. Calculamos el coeficiente de fiabilidad mediante la correlación entre las puntuaciones obtenidas por los sujetos en ambas aplicaciones.

$$r_{X_1X_1'} = r_{x_1x_2} = \frac{N\sum X_1X_2 - \sum X_1 \sum X_2}{\sqrt{\left[N\sum X_1^2 - (\sum X_1)^2\right]\left[N\sum X_2^2 - (\sum X_2)^2\right]}} \qquad [4.27]$$

donde: X_1 y X_2 corresponden, en este caso, a las puntuaciones obtenidas por los sujetos en cada una de las aplicaciones del mismo test.

Como se puede apreciar, el cálculo es idéntico al método de las formas paralelas siendo la única diferencia que en lugar de aplicar dos formas se emplea la misma en dos momentos distintos.

Este método presenta la ventaja de que no se requieren dos ó más formas distintas del mismo test. Con el método test-retest, es el mismo test el que aplicaremos en distintas ocasiones. En el caso de que se pretendan medir rasgos que pueden cambiar en el tiempo hay que extremar las precauciones si tenemos la pretensión de emplear este método ya que se pueden encontrar diferencias en las puntuaciones obtenidas en las dos aplicaciones y no significar falta de estabilidad sino que si realmente los sujetos han variado en el rasgo que se está midiendo, las diferencias pongan de manifiesto ese cambio.

Al igual que el método de las formas paralelas, este método no está exento de inconvenientes que hay que tener presentes. Un primer aspecto a tener en cuenta, es el posible influjo de la memorización de algunos ítems que puede interferir en la segunda aplicación. Un sujeto puede recordar la respuesta que haya dado a ciertos ítems y esto puede provocar un aumento o disminución irreal de su puntuación y, consiguientemente, del valor de la correlación. El efecto de variables de estas características sobre la repetición de un test puede llegar a ser un factor determinante en el valor del coeficiente de fiabilidad.

Un segundo inconveniente a tener en cuenta es el intervalo de tiempo transcurrido entre una aplicación y otra. Es deseable incrementar el tiempo entre aplicaciones para minimizar el efecto de aprendizaje o de memoria pero, al mismo tiempo, un incremento demasiado grande, hace que aumente la posibilidad de que el rasgo que estamos estudiando haya variado debido a la influencia de factores sociales, afectivos o incluso evolutivos propios del sujeto y esto puede incidir en una infraestimación del coeficiente de fiabilidad.

Una última cuestión es la propia actitud del sujeto (Ghiselli, 1981). Un cambio en el grado de cooperación por parte de un sujeto puede provocar, deliberadamente, una puntuación más baja o más alta, que daría como resultado un coeficiente de fiabilidad más bajo o más alto.

Teniendo en cuenta estos aspectos, y si las condiciones de aplicación del test en ambas ocasiones son lo más parecidas posibles, los resultados obtenidos indicarán el grado de estabilidad en las puntuaciones obtenidas. Al coeficiente de fiabilidad así obtenido se le denomina también *coeficiente de estabilidad*.

9. LA FIABILIDAD COMO CONSISTENCIA INTERNA

Existen situaciones en las cuales solamente es posible llevar a cabo una única aplicación de un test; situaciones en las que la aplicación de cualquiera de los dos métodos que acabamos de describir no sea factible, o donde un análisis de la estabilidad o la equivalencia de las medidas no constituya nuestro fin prioritario.

En este apartado presentamos una serie de métodos para estimar la fiabilidad de un test que sólo requieren una aplicación. Unos hacen referencia a la división del test en dos mitades. Otros requieren un análisis de la varianza y covarianza de las respuestas de los sujetos a los ítems. Las diferentes técnicas que presentamos aportan un índice de la *consistencia interna* de las respuestas de los sujetos a los ítems del test en una sola aplicación.

9.1. Métodos basados en la división del test en dos mitades

El método de las dos mitades presenta una ventaja sustancial respecto a los dos métodos explicados anteriormente. Esta ventaja reside en el hecho de que consideramos las puntuaciones obtenidas en una única aplicación de un test, con lo cuál, la estimación de la fiabilidad no se ve afectada por factores como el intervalo de tiempo transcurrido entre una aplicación y otra, la memoria, el aprendizaje, etc., y supone un ahorro de tiempo y esfuerzo al no tener que construir una segunda forma paralela del test, ó tener que realizar una segunda evaluación de los sujetos. Básicamente se trataría de aplicar el test a una muestra de sujetos y, una vez obtenidas las puntuaciones dividir el test en dos mitades, calculando, posteriormente, la correlación entre las puntuaciones obtenidas por los sujetos en ambas partes y aplicar, a continuación, una fórmula de corrección que ya se especificará.

La división del test en dos mitades no es siempre una labor tan sencilla como pueda parecer a primera vista. Las mitades del test deberán ser similares en dificultad y contenido para que la correlación entre las puntuaciones se aproxime al valor máximo. Uno puede cuestionarse si efectivamente las medias, varianzas y el contenido de los ítems son realmente similares o no, y, por lo tanto, si este método es adecuado en todo tipo de situaciones. El hecho de conseguir una igualdad de los valores de la media y la desviación típica es posible con este tipo de agrupamiento, pero como establece Gulliksen (1987) corremos el peligro de agrupar ítems análogos en un solo lado, con lo que pudiera ocurrir que las dos mitades no fueran iguales en cuanto a contenido se refiere. Este aspecto deberá ser cuidadosamente estudiado debido a su importancia.

Son diversas las formas en las que se puede llevar a cabo la división del test en dos mitades pero, ante todo, una característica que habrá que valorar es la forma en que se ha construido el test.

Una primera forma consistiría en dividir el test por la mitad, es decir, considerar los primeros ($n/2$) ítems como una mitad y los últimos ($n/2$) ítems como la segunda mitad. Esta forma de divi-

dir el test puede presentar inconvenientes, puesto que muchos tests están formados por ítems cuya dificultad se va incrementando y, por lo tanto, las dos mitades no serían equivalentes; en el caso de tests con contenidos heterogéneos las dos mitades no serían comparables, y en el caso de tests con un número elevado de ítems hay que tener en cuenta el efecto del cansancio de los sujetos.

Una segunda aproximación al problema consistiría en definir una forma con todos los elementos pares y una segunda forma con todos los elementos impares, con lo cual reducimos significativamente los problemas planteados por la forma anterior.

Una tercera forma de abordar el problema puede ser ordenar los ítems en función de su grado de dificultad, calculando para ello el índice de dificultad de cada ítem, y subdividirlos en pares e impares.

Una cuarta forma, aunque no muy recomendable por razones obvias, podría consistir en la asignación de los ítems al azar a una mitad o a otra.

Normalmente, dado que cuando los ítems del test son de dificultad creciente aparecen ya ordenados a lo largo del test, la forma más utilizada en la división del test en dos mitades, es asignar a una de las mitades los elementos pares y a la otra los impares.

Cuando se utiliza el método de las dos mitades la fiabilidad se puede estimar aplicando cualquiera de las siguientes fórmulas: Spearman-Brown, Rulon, Guttman-Flanagan.

9.1.1. Spearman-Brown

La ecuación de Spearman-Brown, constituye una de las formas más utilizadas para estimar la fiabilidad de un test por el método de las dos mitades. Está basada en la relación existente entre la longitud de un test y el coeficiente de fiabilidad.

En primer lugar aplicamos el test a una muestra de sujetos. Una vez aplicado el test, dividimos éste en dos mitades que han de ser paralelas. Por lo tanto, para ver si la aplicación de este método es la correcta, habría que comprobar los supuestos de paralelismo comentados anteriormente. A continuación calculamos la correlación entre las puntuaciones obtenidas por los sujetos en ambas partes. La correlación calculada correspondería al coeficiente de fiabilidad de cada una de las mitades del test, pero como lo que queremos es calcular la fiabilidad del test completo, para ello aplicamos la ecuación de Spearman-Brown para el caso de longitud doble:

$$R_{xx} = \frac{2r_{xx}}{1+r_{xx}} \qquad [4.28]$$

donde:

R_{xx} = coeficiente de fiabilidad del test.

r_{xx} = coeficiente de fiabilidad de cada una de las mitades.

EJEMPLO:

Hemos aplicado un test de aptitud numérica compuesto de 20 ítems a una muestra de 6 sujetos. Los resultados que se presentan a continuación corresponden a las puntuaciones que dichos sujetos obtuvieron en los ítems pares (X_1) e impares (X_2). Calcular el coeficiente de fiabilidad suponiendo que las dos mitades del test sean paralelas.

SUJETOS	X_1	X_2	X_1^2	X_2^2	X_1X_2
1	8	4	64	16	32
2	7	7	49	49	49
3	8	6	64	36	48
4	5	4	25	16	20
5	8	7	64	49	56
6	6	6	36	36	36
Total	42	34	302	202	241

$$r_{x_1x_2} = \frac{N\sum X_1X_2 - \sum X_1 \sum X_2}{\sqrt{\left[N\sum X_1^2 - (\sum X_1)^2\right]\left[N\sum X_2^2 - (\sum X_2)^2\right]}}$$

$$r_{x_1x_2} = \frac{6\times 241 - 42\times 34}{\sqrt{(6\times 302 - 42^2)(6\times 202 - 34^2)}} = \frac{1.446 - 1.428}{\sqrt{48\times 56}} = 0{,}35$$

$$R_{xx} = \frac{2r_{xx}}{1+r_{xx}} = \frac{2\times 0{,}35}{1+0{,}35} = \frac{0{,}70}{1{,}35} = 0{,}52$$

El coeficiente de fiabilidad de cada una de las mitades es 0,35 pero el del test total es 0,52. Se trata de un coeficiente medio ya que el valor máximo es la unidad. Hemos asumido que las dos mitades son paralelas dado que se trata de un ejemplo, no obstante para aplicar este procedimiento de forma estricta habría que haber hecho previamente la comprobación analizando, por ejemplo, la igualdad de las medias de ambas mitades y la igualdad de los errores típicos de medida.

9.1.2. Rulon

La fórmula de Rulon (1939) para la estimación de la fiabilidad de un test según el método de dos mitades se utiliza cuando, aún no siendo las dos mitades definidas estrictamente paralelas, podemos considerarlas τ-equivalentes (tau-equivalentes) o esencialmente τ-equivalentes. Lord y Novick (1968) definen los tests τ equivalentes como aquellos en los que las puntuaciones verdaderas de los sujetos de una muestra son iguales en ambas formas, pero las varianzas de error no tienen porqué ser iguales, y definen los tests esencialmente (τ) tau-equivalentes como aquellos en los que la puntuación verdadera de cada sujeto en uno de los tests es igual a la del otro más una constante. Tanto en una situación como en otra se asume el cumplimiento del supuesto de igualdad de las varianzas verdaderas de ambas mitades.

Calculados los valores de las puntuaciones en los ítems pares e impares, se calcula la diferencia entre ellas y, a continuación, su varianza (varianza de la diferencia entre las puntuaciones).

$$r_{xx} = 1 - \frac{S_d^2}{S_x^2} \qquad [4.29]$$

donde:

d = diferencias entre las puntuaciones de los elementos pares e impares de cada uno de los sujetos.

$S_d^2 = S_{p-i}^2$ = varianza de la diferencia entre las puntuaciones pares e impares.

S_x^2 = varianza de las puntuaciones empíricas de los sujetos.

EJEMPLO:

Hemos aplicado un test de fluidez verbal compuesto de 6 ítems a 6 sujetos. A continuación se presentan las puntuaciones empíricas obtenidas por los sujetos en el test total, así como las obtenidas en los elementos pares e impares. Calcular el coeficiente de fiabilidad del test.

SUJETOS	X	P	I	(P–I) = d
A	4	3	1	2
B	1	1	0	1
C	6	3	3	0
D	2	1	1	0
E	3	1	2	-1
F	5	2	3	-1

$$\bar{X} = \frac{4+1+6+2+3+5}{6} = 3,5$$

$$S_x^2 = \frac{4^2+1^2+6^2+2^2+3^2+5^2}{6} - (3,5)^2 = 15,17 - 12,25 = 2,92$$

$$\bar{X}_d = 0,17 \qquad S_d^2 = \frac{4+1+1+1}{6} - (0,17)^2 = 1,14$$

$$r_{xx} = 1 - \frac{S_d^2}{S_x^2} = 1 - \frac{1,14}{2,92} = 0,61$$

Se ha obtenido un coeficiente de fiabilidad medio.

9.1.3. Guttman-Flanagan

Flanagan (1937) y Guttman (1945), de forma independiente llegaron a una fórmula equivalente a la de Rulon y que presenta una mayor sencillez de aplicación. La fórmula de Guttman-Flanagan viene determinada por la siguiente expresión:

$$R_{xx} = 2\left(1 - \frac{S_p^2 + S_i^2}{S_x^2}\right) \qquad [4.30]$$

donde:

S_p^2 y S_i^2 = varianzas de las puntuaciones en los ítems pares e impares respectivamente.

S_X^2 = varianza empírica del test total.

Tanto la ecuación de Rulon como la ecuación de Guttman-Flanagan proporcionan el mismo valor de la fiabilidad por ser expresiones equivalentes. Dicha relación aparece recogida al final del tema en el Apéndice.

EJEMPLO:

Con los datos del ejercicio anterior, calcular el coeficiente de fiabilidad utilizando la fórmula de Guttman-Flanagan.

$$\bar{X}_p = 1,83 \quad \bar{X}_i = 1,66$$
$$S_p^2 = 0,81 \quad S_i^2 = 1,21$$

$$S_p^2 = \frac{3^2+1^2+3^2+1^2+1^2+2^2}{6} - (1,83)^2 = 4,16 - 3,35 = 0,81$$

$$S_i^2 = \frac{1^2+0^2+3^2+1^2+2^2+3^2}{6} - (1,67)^2 = 4 - 2,79 = 1,21$$

$$R_{xx} = 2\left(1 - \frac{S_1^2 + S_2^2}{S_x^2}\right) = 2\left(1 - \frac{0,81+1,21}{2,92}\right) = 0,61$$

como puede observarse el resultado es el mismo que el obtenido mediante la fórmula de Rulon.

9.2. Métodos basados en la covariación entre los ítems

Al hablar de la fiabilidad como consistencia interna hemos hecho alusión a dos formas de abordar el tema. Una forma es la basada en la división del test en dos mitades. La segunda forma requiere un análisis de la varianza y covarianza de las respuestas de los sujetos a los ítems. De esta forma, el coeficiente obtenido proporciona una estimación de la consistencia interna de los ítems del test. En el presente apartado haremos referencia a algunos de los métodos más frecuentes para estimar la fiabilidad de un test bajo estas condiciones como son el coeficiente alpha de Cronbach (1951), ó los coeficientes KR20 y KR21 de Kuder-Richardson (1937). Tanto KR20 como KR21 pueden ser considerados como casos particulares del coeficiente «alpha» de Cronbach en el caso de que los ítems que forman el test sean dicotómicos.

9.2.1. Coeficiente alfa (α) de Cronbach

El coeficiente de Cronbach (1951) constituye un indicador de la consistencia interna del test. Este coeficiente expresa la fiabilidad del test en función del número de ítems y de la proporción de la varianza total del test debida a la covariación entre los ítems. Cuanto más covarien los ítems entre sí mayor será la fiabilidad del test.

La ecuación general del coeficiente «alfa» viene expresada como:

$$\alpha = \frac{n}{n-1}\left(\frac{\sum_{j\neq K}^{n}\sum \mathrm{cov}(jk)}{S_x^2}\right) = \left[\frac{n(r_1)}{1+(n-1)r_1}\right] = \frac{n}{n-1}\left(\frac{S_x^2 - \sum S_j^2}{S_x^2}\right) = \frac{n}{n-1}\left(1 - \frac{\sum S_j^2}{S_x^2}\right) \qquad [4.31]$$

donde:

n = número de elementos del test.

ΣS_j^2 = suma de las varianzas de los elementos del test.

$\Sigma\Sigma$cov (jk) = suma de las covarianzas de los ítems.

S_x^2 = varianza de las puntuaciones en el test.

r_1 = cociente entre la covarianza media de los ítems y su varianza media.

EJEMPLO 1:

Hemos aplicado un test de percepción visual a una muestra de seis sujetos. En la tabla adjunta se presentan las puntuaciones que los sujetos obtuvieron en cada uno de los cinco ítems que forman el test. Se desea saber el valor del coeficiente de fiabilidad del test.

SUJETOS	1	2	3	4	5
A	3	4	3	3	4
B	2	3	2	4	4
C	4	2	2	3	3
D	2	1	1	2	1
E	1	1	1	2	1
F	0	0	1	1	1

$$\bar{X}_1 = 2; \qquad S_1^2 = \frac{3^2+2^2+4^2+2^2+1^2+0^2}{6} - (2)^2 = 1,67$$

$$\bar{X}_2 = 1,83; \quad S_2^2 = \frac{4^2+3^2+2^2+1^2+1^2+0^2}{6} - (1,83)^2 = 1,82$$

$$\bar{X}_3 = 1,67; \quad S_3^2 = \frac{3^2+2^2+2^2+1^2+1^2+1^2}{6} - (1,67)^2 = 0,54$$

$$\bar{X}_4 = 2,5; \quad S_4^2 = \frac{3^2+4^2+3^2+2^2+2^2+1^2}{6} - (2,5)^2 = 0,92$$

$$\bar{X}_5 = 2,33; \quad S_5^2 = \frac{4^2+4^2+3^2+1^2+1^2+1^2}{6} - (2,33)^2 = 1,90$$

$$S_x^2 = \frac{17^2+15^2+14^2+7^2+6^2+3^2}{6} - (10,33)^2 = 27,29$$

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum S_j^2}{S_x^2}\right) = \frac{5}{4}\left(1 - \frac{1,67+1,82+0,54+0,92+1,90}{27,29}\right) = 0,94$$

EJEMPLO 2:

Siendo la covarianza media entre todos los elementos de un test igual a 0,25, averiguar el coeficiente de fiabilidad del test sabiendo que está compuesto por 10 ítems y que la varianza empírica es igual a 40 puntos.

Para resolver el problema hay que partir de que la varianza de una variable compuesta, suma de otras variables, es igual a la suma de las varianzas de todas las variables más la de las covarianzas, o bien a la suma de las varianzas más los n(n-1) términos de covarianza media:

$$S_x^2 = \sum_{j=1}^{n} S_j^2 + n(n-1)\overline{r_{jk}S_j^2}; \text{ despejamos } \sum_{j=1}^{n} S_j^2$$

$$\sum_{j=1}^{n} S_j^2 = 40 - 10 \cdot 9 \cdot 0,25 = 17,5 \text{ y calculamos } \alpha\text{: } \alpha = \frac{10}{9}\left(1 - \frac{17,5}{40}\right) = 0,62$$

9.2.1.1. Estimador insesgado de α

El estimador insesgado de α propuesto por Feldt, Woodruff y Salih (1987) se expresa como:

$$\bar{\alpha} = \frac{(N-3)\hat{\alpha}+2}{N-1} \qquad [4.32]$$

donde:

$\bar{\alpha}$ = estimador insesgado.

$\hat{\alpha}$ = valor de alpha de Cronbach.

N = número de sujetos de la muestra.

A medida que aumenta el número de sujetos de la muestra, el valor del α encontrado y el valor del estimador insesgado se aproximan, siendo iguales cuando $N \to \infty$. En la práctica, a partir de 100 sujetos, se pueden considerar insignificantes las diferencias encontradas. Es decir:

$$\bar{\alpha} \Rightarrow \hat{\alpha}, \text{ cuando } N \to \infty \qquad [4.33]$$

Supongamos que en una muestra de 150 sujetos se les ha aplicado un test y se ha obtenido un valor de α = 0,75.

$$\bar{\alpha} = \frac{(150-3)0,75+2}{150-1} = 0,753$$

Como se puede apreciar, a partir de 100 sujetos la diferencia encontrada entre ambos estimadores es insignificante. Si por el contrario tuviéramos una muestra de 20 sujetos, las diferencias serían mayores.

$$\bar{\alpha} = \frac{(20-3)0,75+2}{20-1} = 0,78$$

9.2.1.2. El coeficiente α como límite inferior del coeficiente de fiabilidad

El coeficiente α puede ser considerado como una estimación del límite inferior del coeficiente de fiabilidad de un test, siendo su valor menor o igual que el coeficiente de correlación $r_{xx'}$ (Guttman, 1945):

$$\alpha \leq r_{xx} \qquad [4.34]$$

El lector interesado puede encontrar una demostración de dicha relación en Muñiz (1998).

El coeficiente α es igual al coeficiente de fiabilidad, $r_{XX'}$ cuando los ítems del test sean paralelos y, por tanto, satisfagan las condiciones de paralelismo que hemos formulado con anterioridad.

Otro estimador del límite inferior del coeficiente de fiabilidad es el coeficiente δ (delta) propuesto por Guttman (1945):

$$\delta_3 = 1 - \left(\sum_{j=1}^{n} \frac{S_j^2}{S_x^2} \right) + \frac{\sqrt{\frac{n}{(n-1)} \sum\sum \text{cov}(j,k)}}{S_x^2} \qquad [4.35]$$

donde:

n = número de elementos del test.

S_j^2 = varianza del elemento j del test.

S_x^2 = varianza del test total.

$\sum\sum \text{cov}(j,k) = S_x^2 - \sum_{j=1}^{n} S_j^2$ = suma de las covarianzas de los ítems

9.2.1.3. Inferencias sobre α

Como acabamos de ver, el coeficiente α nos proporciona una estimación de la fiabilidad de un test basada en la consistencia interna del mismo. En ocasiones queremos ir mas allá, y nos planteamos cuestiones como, por ejemplo, si existe una diferencia significativa entre el valor del coeficiente alfa obtenido en dos o más muestras independientes; si alfa puede tomar un valor concreto en la población; si la diferencia entre dos ó más valores distintos de alfa para una misma muestra de sujetos, es significativa o no; etc. Estos problemas referidos a las inferencias acerca del coeficiente alfa, dieron lugar, a principios de los años 60 del siglo veinte, al desarrollo de la teoría muestral para el coeficiente alfa. Kristof (1963) y Feldt (1965), de forma independiente, derivaron un estadístico de contraste del coeficiente alfa, que se distribuye según una distribución F de Snedecor, a partir del cuál se puede determinar un intervalo confidencial para el valor de α en la población.

Feldt (1969) deriva el estadístico «W» para el caso de que se quieran contrastar dos valores de alfa obtenidos en muestras independientes. Dicho método fue ampliado a «n» muestras independientes a partir del estadístico «UX_1» postulado por Hakstian y Whalen (1976). Feldt (1980) desarrolló un estadístico de contraste para dos valores de alfa obtenidos en la misma muestra y, Woodruff y Feldt (1986) ampliaron esta metodología al caso de «n» coeficientes obtenidos en la misma muestra.

a) Inferencias para un solo valor de α

Cuando estamos interesados en saber si el coeficiente alpha puede tomar un determinado valor en la población o, entre qué valores se encuentra el coeficiente α en la población, podemos aplicar el estadístico propuesto por Kristof (1963) y Feldt (1965) independientemente. Es decir, una vez que hayamos obtenido un determinado valor de alfa en una muestra de sujetos, podemos plantearnos la hipótesis de si el valor obtenido es compatible con el hecho de que α tome un determinado valor en la población. El estadístico de contraste propuesto puede expresarse como:

$$F = \frac{1-\alpha}{1-\hat{\alpha}} \qquad [4.36]$$

donde:

F = se distribuye con $(N-1)$ y $(n-1)(N-1)$ grados de libertad.

α = valor de alpha propuesto por hipótesis para la población.

$\hat{\alpha}$ = valor de alpha obtenido en la muestra.

N = número de sujetos.

n = número de ítems.

El siguiente ejemplo clarifica las dos cuestiones a las que hemos hecho referencia.

EJEMPLO:

Supongamos que hemos aplicado un test de percepción espacial compuesto de 35 ítems a una muestra de 60 alumnos de 1º de Bachillerato, y que hemos obtenido un $\alpha = 0{,}83$. Deseamos saber, en primer lugar, si dicho coeficiente es estadísticamente significativo y, en segundo lugar, entre qué valores se encontrará el coeficiente alfa en la población (nivel de confianza del 95%).

La primera cuestión se refiere a si el valor del coeficiente α obtenido es estadísticamente significativo o no. La hipótesis nula que se plantea es H_0: $\alpha = 0$, y como hipótesis alternativa H_1: $\alpha \neq 0$.

[1] Obsérvese que la F que aparece en el denominador tiene invertidos los grados de libertad.

$$F = \frac{1-\alpha}{1-\check{\alpha}} = \frac{1-0}{1-0,83} = 5,88$$

$$gl = (N-1),(n-1)(N-1) = (59,\ 2006)$$

$$F_{0,975} = 1,39; \qquad F_{0,025} = 0,67^{(*)}$$

$$(*)\ F_{0,025,59,2006} = \frac{1}{F_{0,975,2006,59}} = \frac{1}{1,48} = 0,67$$

Puesto que el valor de F obtenido no se encuentra dentro del intervalo establecido se rechaza la H_0 y se puede concluir que el coeficiente alfa es estadísticamente significativo.

La segunda cuestión que nos planteamos es cómo determinar los valores entre los que se encontrará el coeficiente α de la población.

$$\frac{1-\alpha}{1-0,83} \leq 1,39; \qquad \alpha \geq 1-1,39(1-0,83); \qquad \alpha \geq 0,76$$

$$\frac{1-\alpha}{1-0,83} \geq 0,67; \qquad \alpha \leq 1-0,67(1-0,83); \qquad \alpha \leq 0,89$$

$$0,76 \leq \alpha \leq 0,89$$

Al nivel de confianza del 95%, el valor de coeficiente α está comprendido entre 0,76 y 0,89. Por tanto, el valor planteado por la H_0 no está incluido en el intervalo.

b) Inferencias sobre alfa para muestras independientes

Analizaremos dos situaciones: dos muestras independientes y «*K*» muestras independientes.

b.1) Dos muestras independientes

Para el caso de dos muestras independientes, Feldt (1969) propuso el estadístico de contraste W que permite comprobar la *Ho:* $\alpha_1 = \alpha_2$. H_1: $\alpha_1 \neq \alpha_2$

$$W = \frac{1-\hat{\alpha}_1}{1-\hat{\alpha}_2} \qquad [4.37]$$

donde:

W = se distribuye según F con $(N_1 - 1)$ y $(N_2 - 1)$ grados de libertad.

$\hat{\alpha}_1$ y $\hat{\alpha}_2$ = valores del coeficiente alfa en cada una de las muestras.

N_1 y N_2 = número de sujetos de cada muestra.

EJEMPLO:

Hemos aplicado un test de razonamiento, a una muestra de 121 sujetos, obteniendo un valor de alfa igual a 0,55. Se aplicó el mismo test a otra muestra de 61 sujetos, obteniéndose un valor de alfa igual a 0,62. Queremos saber si existen diferencias estadísticamente significativas entre los valores de ambos coeficientes (N.C. 95%).

$$H_0 : \alpha_1 = \alpha_2$$

$$H_1 = \alpha_1 \neq \alpha_2$$

$$W = \frac{1-0,55}{1-0,62} = 1,18$$

$$F_{0,975(120,60)} = 1,58$$

$$F_{0,025(120,60)} = 0,65$$

Podemos afirmar, al N.C. 95%, que la diferencia entre ambos coeficientes no es estadísticamente significativa puesto que el valor $W = 1,18$ se encuentra entre los valores encontrados.

b.2) «K» muestras independientes

Woodruff y Feldt (1986) ampliaron el estudio de Feldt (1969) para el caso de «*K*» coeficientes obtenidos en *K* muestras independientes. Bajo la condición de muestras independientes han derivado el estadístico de contraste UX_1:

$$UX_1 = \frac{\sum^{k}\left[\left(1-\hat{\alpha}_i\right)^{-1/3} - \bar{u}\right]^2}{\bar{S}^2} \qquad [4.38]$$

donde:

UX_1 = se distribuye aproximadamente como χ^2 con *K*–1 grados de libertad.

K = número de muestras o coeficientes.

$\hat{\alpha}_i$ = valor del coeficiente alfa para cada muestra.

$\bar{u}$ = media de los coeficientes transformados.

$$\bar{u} = \sum_{i=1}^{k} \frac{(1-\hat{\alpha}_i)^{-1/3}}{K}$$

$\bar{S}^2$ = media aritmética de las varianzas de cada muestra.

$$\bar{S}^2 = \sum_{i=1}^{k} \frac{S_i^2}{K}$$

siendo:

$$S_i^2 = \frac{2}{9(\tilde{N}_i - 1)(1-\hat{\alpha}_i)^{2/3}}$$

y

$$\tilde{N}_i = \frac{N_i(n_i - 1)}{n_i + 1}$$

donde:

N_i = número de sujetos en cada muestra.

n_i = número de ítems en cada test.

EJEMPLO:

Se ha aplicado un test compuesto por 50 ítems a tres muestras independientes de 25, 40 y 50 sujetos. Para cada una de estas muestras se obtuvieron los siguientes valores de alfa: $\alpha_1 = 0{,}55$, $\alpha_2 = 0{,}70$ y $\alpha_3 = 0{,}75$. Deseamos saber si existen diferencias estadísticamente significativas para los valores de alfa obtenidos (N.C. 95%).

$$H_0: \quad \alpha_1 = \alpha_2 = \alpha_3$$
$$H_1: \quad \alpha_1 \neq \alpha_2 \neq \alpha_3$$

$$\bar{u} = \frac{(1-0{,}55)^{-1/3}}{3} + \frac{(1-0{,}70)^{-1/3}}{3} + \frac{(1-0{,}75)^{-1/3}}{3} = 1{,}457$$

$$\tilde{N}_1 = \frac{25(50-1)}{50+1} = 24{,}02; \qquad S_1^2 = \frac{2}{9(24{,}02-1)(1-0{,}55)^{2/3}} = 0{,}016$$

$$\tilde{N}_2 = \frac{40(50-1)}{50+1} = 38{,}43; \qquad S_2^2 = \frac{2}{9(38{,}43-1)(1-0{,}70)^{2/3}} = 0{,}013$$

$$\tilde{N}_3 = \frac{50(50-1)}{50+1} = 48{,}04; \qquad S_3^2 = \frac{2}{9(48{,}04-1)(1-0{,}75)^{2/3}} = 0{,}011$$

$$\bar{S}^2 = \frac{0{,}016 + 0{,}013 + 0{,}011}{3} = 0{,}013$$

$$UX_1 = \frac{\left[(1-0{,}55)^{-1/3} - 1{,}457\right]^2}{0{,}013} + \frac{\left[(1-0{,}70)^{-1/3} - 1{,}457\right]^2}{0{,}013} +$$

$$+ \frac{\left[(1-0{,}75)^{-1/3} - 1{,}457\right]^2}{0{,}013} = 1{,}778 + 0{,}104 + 1{,}308 = 3{,}19$$

$$gl.(n-1) = 2; \qquad \chi^2_{0{,}975,2} = 7{,}38 \qquad \chi^2_{0{,}025,2} = 0{,}05$$

Podemos afirmar, al N.C. 95%, que no existen diferencias estadísticamente significativas entre los distintos valores de alpha.

c) Inferencias sobre alfa para muestras dependientes

En algunos diseños experimentales es posible administrar distintas pruebas a la misma muestra de sujetos. En estas situaciones los coeficientes obtenidos son dependientes y no podemos emplear ninguno de los dos contrastes que acabamos de estudiar.

Los primeros estudios llevados a cabo para establecer un estadístico de contraste que nos permitiera ver si existen diferencias significativas entre dos coeficientes obtenidos en la misma muestra, fueron llevados a cabo por Feldt (1980) y, posteriormente desarrollados para «*K*» muestras por Woodruff y Feldt (1986).

c.1) Dos muestras dependientes

Feldt (1980, 1987) propuso el empleo del estadístico de contraste «*t*» para dos valores de alfa obtenidos a partir de una misma muestra de sujetos. Feldt recomienda el empleo de este estadístico cuando N · n ≤ 1.000, siendo *N* igual al número de sujetos y *n* el número de ítems. El estadístico se expresa como:

$$t = \frac{|\hat{\alpha}_1 - \hat{\alpha}_2|\sqrt{N-2}}{\sqrt{\left[4(1-\hat{\alpha}_1)(1-\hat{\alpha}_2)(1-r^2_{x_1x_2})\right]}} \qquad [4.39]$$

donde:

t = se distribuye según una distribución t de Student con $(N-2)$ grados de libertad.

$\hat{\alpha}_1$ y $\hat{\alpha}_2$ = valores del coeficiente alfa en los dos tests.

N = número de sujetos en la muestra.

$r^2_{x_1x_2}$ = correlación al cuadrado entre las puntuaciones de los sujetos en los dos tests.

EJEMPLO:

Aplicamos dos tests de percepción visual a una muestra de 125 sujetos. La correlación entre las puntuaciones de ambos tests es igual a 0,70. Los valores del coeficiente alfa fueron, respectivamente: 0,75 y 0,84. Queremos saber si la diferencia existente entre ambos valores es estadísticamente significativa o no (N.C. 95%).

$$H_0 = \alpha_1 = \alpha_2$$

$$H_1 = \alpha_1 \neq \alpha_2$$

$$t = \frac{|0,84-0,75|\sqrt{(125-2)}}{\sqrt{\left[4(1-0,84)(1-0,75)(1-0,70^2)\right]}} = 3,50$$

$$t_{(N-2)} = t_{123} = 1,96$$

Se rechaza la hipótesis nula y podemos establecer que la diferencia entre los coeficientes es estadísticamente significativa.

c.2) «K» muestras dependientes

Para el caso de «*K*» muestras, Woodruff y Feldt (1986) presentaron una serie de estadísticos de contraste entre los que cabe resaltar, por su sencillez de aplicación y gran precisión, el estadístico UX_2. Si bien es cierto que el cálculo no presenta grandes dificultades, no es menos cierto que es algo laborioso, por lo que han sido muchos los investigadores que han intentado desarrollar programas para realizar los cálculos de manera informatizada. Entre ellos, cabe destacar el desarrollado por Lautenschiager (2013).

$$UX_2 = \frac{\sum_{i=1}^{k}\left[(1-\hat{\alpha}_i)^{-1/3} - \bar{u}\right]^2}{\bar{S}^2 - \bar{C}} \qquad [4.40]$$

donde:

UX_2 = se distribuye aproximadamente igual a χ^2 con $(K-1)$ grados de libertad.

K = número de muestras o coeficientes.

N = número de sujetos de la muestra.

$\hat{\alpha}_i$ = valor de los coeficientes alfa.

$\bar{u}$ = media de los coeficientes transformados.

$$\bar{u} = \sum_{i=1}^{k}\frac{[1]}{K(1-\hat{\alpha}_i)^{1/3}}$$

$\bar{S}^2$ = media aritmética de las varianzas de cada muestra.

$$\bar{S}^2 = \sum_{i=1}^{k}\frac{S_i^2}{k}$$

donde:

$$S_i^2 = \frac{2}{9(\tilde{N}-1)(1-\hat{\alpha}_i)^{2/3}}$$

siendo:

$$\tilde{N} = \frac{N(\bar{n}-1)}{\bar{n}+1}$$

y

$$\tilde{n} = \frac{K}{\sum_{i=1}^{k} \frac{1}{n_i}} \text{ (media armónica de las longitudes de los tests)}$$

donde:

n_i = número de ítems de cada test.

$\overline{C}$ = media de las covarianzas S_{jk}.

$$C = \frac{2r_{jk}^2}{9\left(\tilde{N}-1\right)\left(1-\hat{\alpha}_j\right)^{1/3}\left(1-\hat{\alpha}_k\right)^{1/3}}$$

EJEMPLO:

Se aplicaron 3 versiones de un cuestionario de ansiedad a una muestra de 100 sujetos. Los cuestionarios estaban compuestos de $A = 50$, $B = 60$ y $C = 65$ ítems respectivamente. Los coeficientes alfa obtenidos fueron: $\alpha_A = 0{,}60$, $\alpha_B = 0{,}70$ y $\alpha_C = 0{,}74$. Las correlaciones entre las puntuaciones de los sujetos fueron: $r_{AB} = 0{,}50$; $r_{AC} = 0{,}58$ y $r_{BC} = 0{,}59$. Calcular, al N.C. 95%, si existen diferencias significativas entre los valores de los coeficientes α obtenidos:

$$H_0 = \alpha_A = \alpha_B = \alpha_C$$

$$H_1 = \alpha_A \neq \alpha_B \neq \alpha_C$$

$$\overline{u} = \frac{1}{3(1-0{,}60)^{1/3}} + \frac{1}{3(1-0{,}70)^{1/3}} + \frac{1}{3(1-0{,}74)^{1/3}} = 0{,}45 + 0{,}50 + 0{,}52 = 1{,}47$$

$$\tilde{n} = \frac{3}{\frac{1}{50}+\frac{1}{60}+\frac{1}{65}} = \frac{3}{0{,}02+0{,}016+0{,}015} = \frac{3}{0{,}051} = 58{,}82$$

$$\tilde{N} = \frac{100(58{,}82-1)}{58{,}82+1} = 96{,}65$$

$$S_A^2 = \frac{2}{9(96{,}65-1)(1-0{,}60)^{2/3}} = 0{,}0042$$

$$S_B^2 = \frac{2}{9(96{,}65-1)(1-0{,}70)^{2/3}} = 0{,}0052$$

$$S_C^2 = \frac{2}{9(96{,}65-1)(1-0{,}74)^{2/3}} = 0{,}0057$$

$$\overline{S}^2 = \sum_{i=1}^{K} \frac{S_i^2}{k} = \frac{0{,}0042+0{,}0052+0{,}0057}{3} = 0{,}0050$$

$$C_{AB} = \frac{2(0{,}50)^2}{9(96{,}65-1)(1-0{,}60)^{1/3}(1-0{,}70)^{1/3}} = 0{,}0011$$

$$C_{AC} = \frac{2(0{,}58)^2}{9(96{,}65-1)(1-0{,}60)^{1/3}(1-0{,}74)^{1/3}} = 0{,}0016$$

$$C_{BC} = \frac{2(0{,}59)^2}{9(96{,}65-1)(1-0{,}70)^{1/3}(1-0{,}74)^{1/3}} = 0{,}0019$$

$$\overline{C} = \frac{0{,}0011+0{,}0016+0{,}0019}{\frac{3(3-1)}{2}} = 0{,}0015$$

$$UX_2 = \frac{\left[(1-0{,}60)^{-1/3}-1{,}47\right]^2}{0{,}0035} + \frac{\left[(1-0{,}70)^{-1/3}-1{,}47\right]^2}{0{,}0035} + \frac{\left[(1-0{,}74)^{-1/3}-1{,}47\right]^2}{0{,}0035} =$$

$$= 3{,}63 + 0{,}16 + 2{,}68 = 6{,}47$$

$$g.l.(n-1) = 2; \quad \begin{matrix} \chi^2_{0{,}975,2} = 7{,}38 \\ \chi^2_{0{,}025,2} = 0{,}05 \end{matrix}$$

Por lo tanto, el valor obtenido queda dentro del intervalo, y por lo tanto se acepta la hipótesis nula ya que no hay diferencias estadísticamente significativas.

9.2.2. Casos particulares del coeficiente α

En este punto hacemos referencia a la estimación de la fiabilidad de un test en el caso de que los ítems que lo componen sean dicotómicos, para lo cual haremos referencia a los estudios de

Kuder y Richardson (1937, 1939). Las ecuaciones de Kuder-Richardson (1937) representan un caso particular del coeficiente «alpha» de Cronbach, en el supuesto de que los ítems sean dicotómicos. Esta estimación es una función del número de ítems y sus intercorrelaciones. Cuanto mayor sea el número de ítems, y cuanto mayor sea el valor de sus covarianzas, mayor será su consistencia interna, y mayor será la fiabilidad.

Teniendo en cuenta que la ecuación de Kuder-Richardson se basa en que los elementos del test son dicotómicos, éstos vendrán puntuados con un 1, en el caso de acierto (o de respuesta favorable en el caso de que se traten de medir variables no cognitivas) y, con un 0, en el caso de fallo (o respuesta desfavorable en el caso de variables no cognitivas).

Como ya se ha visto, el coeficiente «alpha» puede expresarse:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum S_j^2}{S_x^2}\right)$$

Sabemos, por otra parte, que la varianza de una variable dicotómica cualquiera, «*h*», con proporción de aciertos $\boldsymbol{p_h}$, y proporción de errores $\boldsymbol{q_h}$, siendo $\boldsymbol{q_h = 1 - p_h}$, podemos expresarla en los siguientes términos:

$$S_h^2 = p_h q_h$$

con lo que la ecuación del coeficiente «alpha» que acabamos de ver puede escribirse:

$$KR_{20} = \frac{n}{n-1}\left(1 - \frac{\sum p_h q_h}{S_x^2}\right) \qquad [4.41]$$

donde:

n = número de elementos del test.

p_h = proporción de aciertos en el elemento *h*. $p_h = \frac{f_h}{N}$, igual también a la media del elemento.

q_h = proporción de errores en el elemento *h*. $q_h = 1 - p_h$

$p_h q_h$ = varianza del elemento *h*.

S_x^2 = varianza total del test.

Dicha expresión recibe el nombre de ecuación de Kuder-Richardson20 (KR20).

Si los ítems que componen el test, además de ser dicotómicos, presentan la misma dificultad, podemos aplicar la ecuación de Kuder-Richardson 21 (KR21).

$$KR_{21} = \frac{n}{n-1}\left(1 - \frac{npq}{S_x^2}\right) \qquad [4.42]$$

donde:

n = número de elementos del test.

npq = suma de las varianzas de los elementos. Al ser iguales las varianzas se sustituye el signo sumatorio por «*n*» veces la misma varianza.

S_x^2 = varianza del test.

La expresión anterior se puede simplificar y expresarse en los siguientes términos:

$$KR_{21} = \frac{n}{n-1}\left(1 - \frac{\bar{X} - \frac{\bar{X}^2}{n}}{S_x^2}\right) \qquad [4.43]$$

donde:

n = número de elementos del test.

S_x^2 = varianza del test.

$\bar{X}$ media de las puntuaciones empíricas.

EJEMPLO:

Supongamos un test *(A)* de fluidez verbal y otro test *(B)* de comprensión lectora, cuyas puntuaciones aparecen en las siguientes matrices de datos. El test de fluidez verbal sólo admite dos posibles puntuaciones, 1 y 0. Calcular el valor del coeficiente de fiabilidad de ambos tests.

Test A

Ítems

Sujetos	A	B	C	D	E	F
1	1	1	1	1	1	1
2	1	1	1	0	1	1
3	1	0	1	0	1	1
4	0	1	0	1	0	1
5	0	0	0	0	0	0
6	1	0	0	0	0	0

Test A

Ítems

Sujetos	A	B	C	D	E	F
1	3	4	3	3	4	3
2	2	3	2	4	4	2
3	4	2	2	3	3	4
4	2	1	1	2	1	2
5	1	1	1	2	1	2
6	0	0	1	1	1	1

Medias de los ítems en el Test B:

$$\bar{X}_1 = 2;\quad \bar{X}_2 = 1{,}83;\quad \bar{X}_3 = 1{,}67\ ;\quad \bar{X}_4 = 2{,}5;\quad \bar{X}_5 = 2{,}33\ \ y\ \ \bar{X}_6 = 2{,}33$$

TEST A

$$p_1 = \frac{4}{6} = 0{,}67 \qquad q_1 = 1 - 0{,}67 = 0{,}33 \qquad p_1 q_1 = 0{,}67 \cdot 0{,}33 = 0{,}22$$

$$p_2 = \frac{3}{6} = 0{,}50 \qquad q_2 = 1 - 0{,}50 = 0{,}50 \qquad p_2 q_2 = 0{,}50 \cdot 0{,}50 = 0{,}25$$

$$p_3 = \frac{3}{6} = 0{,}50 \qquad q_2 = 1 - 0{,}50 = 0{,}50 \qquad p_2 q_2 = 0{,}50 \cdot 0{,}50 = 0{,}25$$

$$p_4 = \frac{2}{6} = 0{,}33 \qquad q_4 = 1 - 0{,}33 = 0{,}67 \qquad p_4 q_4 = 0{,}33 \cdot 0{,}67 = 0{,}22$$

$$p_5 = \frac{3}{6} = 0{,}50 \qquad q_5 = 1 - 0{,}50 = 0{,}50 \qquad p_5 q_5 = 0{,}50 \cdot 0{,}50 = 0{,}25$$

$$p_6 = \frac{4}{6} = 0{,}67 \qquad q_6 = 1 - 0{,}67 = 0{,}33 \qquad p_6 q_6 = 0{,}67 \cdot 0{,}33 = 0{,}22$$

$$\overline{X_A} = \frac{19}{6} = 3{,}17$$

$$S_A^2 = 4{,}45$$

$$S_A^2 = \frac{\sum X_A^2}{N} = \left(\bar{X}\right)^2 = \frac{36 + 25 + 16 + 9 + 0 + 1}{6} = 10{,}05 = 4{,}5$$

TEST B

$$S_1^2 = \frac{\sum X^2}{N} - \left(\bar{X}\right)^2 = \frac{9 + 4 + 16 + 4 + 1}{6} - (2)^2 = 1{,}67$$

$$S_2^2 = \frac{16 + 9 + 4 + 1 + 1}{6} - (1{,}83)^2 = 1{,}82$$

$$S_3^2 = \frac{9 + 4 + 4 + 1 + 1 + 1}{6} - (1{,}67)^2 = 0{,}54$$

$$S_4^2 = \frac{9 + 16 + 9 + 4 + 4 + 1}{6} - (2{,}50)^2 = 0{,}92$$

$$S_5^2 = \frac{16 + 16 + 9 + 1 + 1 + 1}{6} - (2{,}33)^2 = 1{,}90$$

$$S_6^2 = \frac{9 + 4 + 16 + 4 + 4 + 1}{6} - (2{,}33)^2 = 0{,}90$$

$$\bar{X}_B = \frac{76}{6} = 12{,}67\ ; \qquad S_B^2 = \frac{1174}{6} - 12{,}67^2 = 35{,}14$$

$$R_{aa} = KR_{20} = \frac{n}{n-1}\left(1 - \frac{\sum p_h q_h}{S_x^2}\right) = \frac{6}{5}\left(1 - \frac{0{,}22 + 0{,}25 + 0{,}25 + 0{,}22 + 0{,}25 + 0{,}22}{4{,}45}\right) = 0{,}82$$

$$R_{bb} = \alpha = \frac{n}{n-1}\left(1 - \frac{\sum S_j^2}{S_x^2}\right) = \frac{6}{5}\left(1 - \frac{1{,}67 + 1{,}82 + 0{,}54 + 0{,}92 + 1{,}90 + 0{,}90}{35{,}14}\right) = 0{,}94$$

En el caso de aplicar KR_{21} con ítems cuya dificultad no es la misma, se obtendrá un valor inferior al de KR_{20}. En el test A, que es el que tiene los ítems dicotómicos el valor encontrado sería:

$$KR_{21} = \frac{n}{n-1}\left(1 - \frac{\bar{X} - \frac{\bar{X}^2}{n}}{S_x^2}\right) = \frac{6}{5}\left(1 - \frac{3{,}17 - \frac{3{,}17^2}{6}}{4{,}45}\right) = 0{,}80$$

Como se puede observar el valor obtenido mediante la fórmula KR_{21} es algo inferior al obtenido mediante la KR_{20}, lo que indica que los ítems del test no tienen la misma dificultad.

9.3. Coeficientes basados en el análisis factorial de los ítems: Theta (θ) y Omega (Ω)

Los coeficientes Theta (θ) de Carmines (Carmines y Zeller, 1979) y Omega (Ω) de Heise y Bohrnstedt (1970) constituyen dos indicadores de la consistencia interna de los ítems de un test y una aproximación al coeficiente alpha. Se trata de dos coeficientes basados en el análisis factorial de los ítems.

El coeficiente θ se puede expresar mediante la siguiente fórmula:

$$\theta = \frac{n}{n-1}\left(1-\frac{1}{\lambda_1}\right) \qquad [4.44]$$

donde:

n = número de ítems del test.

λ_1 = primer autovalor de la matriz factorial; es decir, la varianza explicada por el primer factor antes de la rotación.

El coeficiente θ es además un indicador de la unidimensionalidad de los ítems. Cuanto mayor sea la varianza que explica el primer factor mayor será el valor de theta y, por consiguiente, la intercorrelación entre los ítems, lo que implica que éstos se distribuyan en torno a una sola dimensión.

El coeficiente Ω se puede expresar mediante la siguiente fórmula:

$$\Omega = 1-\frac{\sum_{j=1}^{n} S_j^2 - \sum_{j=1}^{n} S_j^2 h_j^2}{\sum_{j=1}^{n}\sum_{\substack{h=1 \\ j\neq h}}^{n} \mathrm{cov}(X_j, X_h)} \qquad [4.45]$$

donde:

ΣS_j^2 = suma de las varianzas de los ítems.

h_j^2 = comunalidad estimada del ítem j.

$\Sigma\Sigma Cov\ (X_j, X_h)$ = suma de las covarianzas entre los ítems j y h.

Otra forma más sencilla de expresar el coeficiente Ω es en función de las correlaciones entre los ítems:

$$\Omega = 1-\frac{n-\sum h_j^2}{n+2\sum r_{jh}} \qquad [4.46]$$

Donde r_{jh} representa la correlación entre los ítems j y h.

En general, y para los mismos datos, se verifica que $\alpha \leq \theta \leq \Omega$. La igualdad entre los coeficientes se verifica cuando los ítems son paralelos (Carmines y Séller, 1979).

EJEMPLO:

En la siguiente tabla aparecen los valores de la varianza explicada por los cinco factores obtenidos tras someter a un análisis factorial a 5 variables. La suma de las comunalidades es igual a 4.95 y la suma de las correlaciones entre los ítems es igual a 5.1. Calcular el valor de los coeficientes θ y Ω.

Factor	Varianza explicada
1	3,286
2	1,346
3	0,224
4	0,128
5	0,014

$$\theta = \frac{n}{n-1}\left(1-\frac{1}{\lambda_1}\right) = \frac{5}{5-1}\left(1-\frac{1}{3,286}\right) = 0,869$$

$$\Omega = 1-\frac{n-\sum h_j^2}{n+2\sum r_{jh}} = 1-\frac{5-4,95}{5+2\cdot 5,1} = 0,996$$

9.4. El coeficiente beta (β) de Raju

Cronbach (1951) introdujo el coeficiente alfa como una medida de la consistencia interna de un test. En el caso de que un test se divida en varios subtests, con desigual número de ítems, y se quiera estimar la consistencia interna del test total a partir de las puntuaciones totales de los suje-

tos en los subtests, el coeficiente alfa presenta el problema de que proporciona un valor infraestimado de la fiabilidad. El coeficiente β propuesto por Raju (Raju, 1977) permite superar este problema y proporciona una estimación adecuada de la fiabilidad de un test compuesto de varios subtests con distinto número de ítems. Se aplica este coeficiente cuando se desconocen las puntuaciones de los sujetos en los ítems de los distintos subtests. En el caso de conocer los valores de estas puntuaciones es mejor emplear el coeficiente α.

El coeficiente β viene dado por la expresión:

$$\beta = \frac{S_x^2 - \sum_{j=1}^{k} S_j^2}{S_x^2\left(1 - \sum_{j=1}^{k'}\left(\frac{n_j}{N}\right)^2\right)} \qquad [4.47]$$

donde:

k = número de subtests.

S_x^2 = varianza del test.

S_j^2 = varianza de cada subtest.

n_j = número de ítems en cada subtest.

N = número de ítems total de la batería.

EJEMPLO:

Hemos aplicado un test de destreza manual, compuesto de cuatro subtests, a una muestra de 200 empleados de correos. Los subtests están compuestos por $A = 18$, $B = 30$, $C = 45$ y $D = 55$ ítems respectivamente. La varianza total del test es igual a 50 y las varianzas de los respectivos subtests iguales a $S_a^2 = 5$, $S_b^2 = 7$, $S_c^2 = 9$ y $S_d^2 = 11$. Calcular el valor de los coeficientes α y β.

$$\alpha = \frac{4}{4-1}\left(1 - \frac{5+7+9+11}{50}\right) = 0,48$$

$$\beta = \frac{50-(5+7+9+11)}{50[1-(0,015+0,041+0,092+0,138)]} = 0,50$$

Si utilizamos el coeficiente α hay que tener en cuenta que el término «n» no es igual al número de ítems sino al número de subtests que forman el test. En el ejemplo que se presenta el test de destreza manual está compuesto por cuatro subtests, de ahí que «n» sea igual a 4.

En el caso de que los distintos subtests contengan el mismo número de ítems, entonces el coeficiente β es igual al coeficiente α. (Véase Apéndice al final del tema)

10. ESTIMACIÓN DE LA PUNTUACIÓN VERDADERA DE LOS SUJETOS EN EL ATRIBUTO DE INTERÉS

Una vez estudiado el problema de cómo poder calcular la fiabilidad de un test mediante los procedimientos descritos anteriormente, estamos en condiciones de poder abordar el problema de cómo hacer estimaciones acerca del valor de la puntuación verdadera de un sujeto en un test y del error que afecta a las puntuaciones empíricas obtenidas en el mismo. Desgraciadamente no podemos calcular el valor exacto de la puntuación verdadera de un sujeto, pero sí establecer un intervalo confidencial dentro del cual se encontrará dicha puntuación con un determinado nivel de confianza. Dentro de este apartado veremos tres formas de llevar a cabo esta estimación: la primera mediante la desigualdad de Chebyshev, donde no se hace ningún supuesto sobre la distribución de las puntuaciones empíricas o de los errores; la segunda basada en la distribución normal de los errores, asumiendo una distribución normal de los errores de medida y de las puntuaciones empíricas; y, la tercera basada en el modelo de regresión lineal de mínimos cuadrados.

10.1. Estimación mediante la desigualdad de Chebyshev

Si no se hace ningún supuesto sobre la distribución de las puntuaciones empíricas o de los errores, se aplica la desigualdad de Chebyshev, que establece que si μ y σ son la media y la desviación estándar de una variable aleatoria X, entonces para cualquier constante positiva k la probabilidad es al menos que X asumirá un valor dentro de k desviaciones estándar de la media. En términos psicométricos podemos expresar esta desigualdad como:

$$\forall K \qquad P\left\{|X - V| \leq K(S_e)\right\} \geq 1 - \frac{1}{K^2} \qquad [4.48]$$

donde:

K = constante que toma valores positivos

$1 - \frac{1}{K^2}$ = nivel de confianza utilizado.

S_e = error típico de medida.

EJEMPLO:

Habiendo administrado a una muestra de 200 sujetos un test de razonamiento numérico, se obtuvieron los siguientes resultados: $\overline{X} = 52$, $S_x = 7$ y $r_{xx} = 0{,}73$. Estimar la puntuación verdadera de un sujeto que obtuvo en el test una puntuación empírica de 65 puntos. Nivel de confianza del 95%.

$$S_e = S_x\sqrt{1-r_{xx}} = 7\sqrt{1-0{,}73} = 3{,}64$$

$$1-\frac{1}{K^2} = 0{,}95;\ \frac{1}{K^2} = 1-0{,}95 = 0{,}05;\ \frac{1}{0{,}05} = K^2;\ K^2 = 20;\ K = \sqrt{20} = 4{,}47 \approx 4{,}5$$

$$P\{|65-V| \le 3{,}64 \cdot 4{,}5\} \ge 0{,}95$$

$$P\{-16{,}38 \le |V-65| \le 16{,}38\} \ge 0{,}95$$

$$P\{48{,}62 \le V \le 81{,}38\} \ge 0{,}95$$

Por lo tanto, la puntuación verdadera se encontrará entre los valores 48,62 y 81,38. Este es, sin embargo, un intervalo confidencial demasiado amplio que conlleva una estimación vaga. Esta amplitud exagerada del intervalo confidencial puede ser debida, en primer lugar, a un coeficiente de fiabilidad bajo o, en segundo lugar, a que el método de Chebyshev no considera el tipo de distribución de las puntuaciones empíricas.

10.2. Estimación basada en la distribución normal de los errores

Este método asume una distribución normal de los errores de medida (con media 0 y varianza S_e^2) y de las puntuaciones empíricas condicionadas a un determinado valor de V.

Para la determinación del intervalo confidencial dentro del que se encontrará la puntuación verdadera del sujeto seguiremos los siguientes pasos:

1) Se fija un nivel de confianza y se determina el valor Z_c correspondiente buscándolo en la tabla de distribución normal. Por ejemplo, para un nivel de confianza del 95% tendremos un valor Z_c igual a 1,96.
2) Calcular el error típico de medida (S_e).

   $S_e = S_x\sqrt{1-r_{xx}}$, para puntuaciones directas o diferenciales

   $S_{z_e} = \sqrt{1-r_{xx}}$, para puntuaciones típicas
3) Calcular el error de medida máximo ($E_{máx}$) que estamos dispuestos a admitir. Este error de medida se verá afectado también por el nivel de confianza adoptado.

$$E_{máx} = Z_c \cdot S_e$$

4) Calcular el intervalo confidencial en el que se encontrará la puntuación verdadera.

$$IC = X \pm E_{máx}$$

EJEMPLO:

Habiendo administrado a una muestra de 200 sujetos un test de razonamiento numérico, se obtuvieron los siguientes resultados, $\overline{X} = 52$, $S_x = 7$ y $r_{xx} = 0{,}73$. Estimar la puntuación verdadera (en puntuaciones directas, diferenciales y típicas) de un sujeto que obtuvo en el test una puntuación empírica directa de 65 puntos. N.C. 95%.

$$X = 65\ ;\quad x = X - \overline{X} = 65-52 = 13\ ;\quad Z_x = \frac{65-52}{7} = 1{,}86$$

$$N.C.95\% \Rightarrow Z_c = \pm 1{,}96$$

$$S_e = S_x\sqrt{1-r_{xx}} = 7\sqrt{1-0{,}73} = 3{,}64$$

$$E_{máx} = Z_c \cdot S_e = 1{,}96 \cdot 3{,}64 = 7{,}13$$

$$I\,C = X \pm E_{máx} = 65 \pm 7{,}13 \rightarrow \begin{cases} 72{,}13 \\ 57{,}87 \end{cases};\ 57{,}85 \le V \le 72{,}13 \quad \text{(Puntuaciones Directas)}$$

$$I\,C = x \pm E_{máx} = 13 \pm 7{,}13 \rightarrow \begin{cases} 20{,}13 \\ 5{,}87 \end{cases};\ 5{,}87 \le v \le 20{,}13 \quad \text{(Puntuaciones Diferenciales)}$$

$$S_{z_e} = \sqrt{1-r_{xx}} = \sqrt{1-0{,}73} = 0{,}52$$

$$E_{máx} = Z_c \cdot S_{Ze} = 1{,}96 \cdot 0{,}52 = 1{,}02$$

$$I.C = Z_x \pm E_{máx} = 1{,}86 \pm 1{,}02 \rightarrow \begin{cases} 2{,}88 \\ 0{,}84 \end{cases}$$

$$0{,}84 \le Z_v \le 2{,}88 \quad \text{(Puntuaciones Típicas)}$$

Como se puede apreciar, con respecto a la estimación según el procedimiento de Chebyshev, el intervalo confidencial se ha reducido sensiblemente.

La principal ventaja que presenta la utilización de un intervalo confidencial, a pesar de las críticas formuladas por Nunnally (1970), es que clarifica el hecho de que una puntuación empírica está afectada por un cierto error de medida. Es decir, si un test presenta un coeficiente de fiabili-

dad bajo y, consiguientemente, poca precisión de medida, los intervalos confidenciales son muy amplios. A medida que dichos coeficientes van incrementándose, los valores extremos del intervalo se acotan denotando una aproximación a la puntuación verdadera del sujeto (Allen y Yen, 1979; Yela, 1984).

10.3. Estimación basada en el Modelo de Regresión

Así como la correlación entre las puntuaciones verdaderas y los errores de medida es igual a cero ($r_{ve} = 0$), no sucede lo mismo entre la correlación de las puntuaciones empíricas de los sujetos y los errores de medida, puesto que dichas puntuaciones se ven afectadas por un cierto componente de error produciéndose un sesgo. Esta correlación vendrá expresada, como ya hemos visto, como $r_{xe} = \sqrt{1 - r_{xx}}$

La correlación así establecida es siempre igual o mayor de cero. Su valor máximo se alcanzará cuando la fiabilidad del test sea nula ($r_{xx} = 0$) y su valor mínimo se alcanzará cuando la fiabilidad del test sea perfecta ($r_{xx} = 1$). En el primer caso las puntuaciones empíricas coincidirán con los errores y, en el segundo caso, no habrá errores y las puntuaciones empíricas coincidirán con las verdaderas.

En cualquier caso, como esa correlación es siempre positiva, las puntuaciones empíricas son siempre sesgadas y, por lo tanto, es más conveniente establecer el intervalo confidencial no a partir de las puntuaciones empíricas (que son sesgadas) sino a partir de la puntuación verdadera estimada, que podremos calcular mediante el modelo de regresión lineal según el criterio de mínimos cuadrados.

Las ecuaciones de la recta de regresión de **Y** sobre **X** vienen expresadas por las siguientes ecuaciones:

— Puntuaciones Directas: $Y' = (\overline{Y} - r_{xy}\frac{S_y}{S_x}\overline{X}) + r_{xy}\frac{S_y}{S_x}X = r_{xy}\frac{S_y}{S_x}(X - \overline{X}) + \overline{Y}$

— Puntuaciones Diferenciales: $y' = r_{xy}\frac{S_y}{S_x}x$ siendo $x = (X - \overline{X})$

— Puntuaciones Típicas: $Z_{y'} = r_{xy}Z_x$ siendo $Z_x = \frac{X - \overline{X}}{S_x}$

Nota: El lector interesado puede encontrar una explicación más detallada en los textos de *Introducción al Análisis de Datos y Diseños de Investigación*.

A partir de dichas ecuaciones de regresión podemos establecer las ecuaciones correspondientes para estimar el valor de la puntuación verdadera. Dichas ecuaciones vendrán expresadas de la siguiente forma:

1. Ecuación de regresión en puntuaciones directas de **V** sobre **X**.

$$V' = r_{xv}\frac{S_v}{S_x}X + (\overline{V} - r_{xv}\frac{S_v}{S_x}\overline{X}) = r_{xv}\frac{S_v}{S_x}(X - \overline{X}) + \overline{V} \qquad [4.49]$$

Sabemos que, $r_{xv}\frac{S_v}{S_x} = \frac{S_v}{S_x}\frac{S_v}{S_x} = r_{xv}^2 = r_{xx}$ y dado que $\overline{V} = \overline{X}$ podemos establecer que:

$$V' = r_{xx}X + (\overline{X} - r_{xx}\overline{X}) \qquad V' = r_{xx}(X - \overline{X}) + \overline{X} \qquad [4.50]$$

2. Ecuación de regresión en puntuaciones diferenciales.

$v' = r_{vx}\frac{S_v}{S_x}x$, como $r_{xv} = \frac{S_v}{S_x}$ tendremos que: $v' = \frac{S_v}{S_x}\frac{S_v}{S_x}x = \frac{S_v^2}{S_x^2}x = r_{xx} \cdot x$

$$v' = r_{xx} \cdot x \qquad v' = r_{xx}(X - \overline{X}) \qquad [4.51]$$

3. Ecuación de regresión en puntuaciones típicas.

$$Z_{v'} = r_{vx} \cdot Z_x \qquad [4.52]$$

EJEMPLO:

Con los datos del ejemplo anterior, estimar la puntuación verdadera de un sujeto que obtuvo en el test una puntuación empírica de 65 puntos. N.C. 95%

Puntuaciones directas:

$$V' = r_{xx}X + (\overline{X} - r_{xx}\overline{X}) = 0,73 \cdot 65 + (52 - 0,73 \cdot 52) = 47,45 + 14,04 = 61,49$$

Puntuaciones diferenciales:

$$v' = r_{xx} \cdot x = 0,73 \cdot (65 - 52) = 9,49$$

Puntuaciones típicas:

$$Z_{v'} = r_{xv} \cdot Z_x = \sqrt{0,73} \cdot \frac{65 - 52}{7} = 0,85 \cdot 1,86 = 1,58$$

Una vez estimado el valor de la puntuación verdadera se seguirá el esquema general con el fin de establecer el intervalo confidencial en el que se pueda aceptar, a un determinado nivel de confianza, que se encuentra la puntuación verdadera del sujeto. Los pasos a seguir serían los siguientes:

— Adoptar un nivel de confianza y determinar el valor zeta crítico (Z_c).

— Calcular el error típico de estimación S_{VX}. Siendo:

$S_{VX} = S_x\sqrt{1 - r_{xx}}\sqrt{r_{xx}}$ (Puntuaciones directas o diferenciales)

$S_{ZvZx} = \sqrt{1 - r_{xx}}\sqrt{r_{xx}}$ (Puntuaciones típicas)

— Calcular el error máximo de estimación $E_{máx}$. Siendo $E_{máx} = Z_c \cdot S_{VX}$ en puntuaciones directas o diferenciales y $E_{máx} = Z_c \cdot S_{ZvZx}$ en puntuaciones típicas.

— Establecer el intervalo confidencial a partir de la estimación puntual obtenida al aplicar las ecuaciones de regresión.

Dicho intervalo viene expresado por: $V' \pm E_{máx}$, ó $v' \pm E_{máx}$, ó $Z_{v'} \pm E_{máx}$

Para los datos del ejemplo anterior tenemos:

N.C. 95% $\Rightarrow Z_c = \pm 1,96$

$S_{vx} = S_x\sqrt{1 - r_{xx}}\sqrt{r_{xx}} = 7\sqrt{1 - 0,73}\sqrt{0,73} = 3,09$

$E_{máx} = Z_c \cdot S_{VX} = 1,96 \cdot 3,09 = 6,06$

$IC = V' \pm E_{máx} = 61,49 \pm 6,06 \rightarrow \begin{cases} 67,55 \\ 55,43 \end{cases}$ En puntuaciones directas

$I.C. = v' \pm E_{máx} = 9,49 \pm 6,06 \rightarrow \begin{cases} 15,55 \\ 3,43 \end{cases}$ En puntuaciones diferenciales

$S_{ZvZx} = \sqrt{1 - r_{xx}}\sqrt{r_{xx}} = \sqrt{1 - 0,73}\sqrt{0,73} = 0,44$

$E_{máx} = Z_c \cdot S_{ZvZx} = 1,96 \cdot 0,44 = 0,86$

$I.C. = Z_{v'} \pm E_{máx} = 1,58 \pm 0,86 \rightarrow \begin{cases} 2,44 \\ 0,72 \end{cases}$ En puntuaciones típicas

11. FIABILIDAD DE UNA BATERÍA DE TESTS

Se trata de calcular la fiabilidad de la batería en función de los coeficientes de fiabilidad, varianzas y covarianzas de los subtests que la van a conformar.

La fórmula a utilizar en este caso será:

$$r_{tt} = 1 - \frac{\sum S_j^2 - \sum S_j^2 r_{jj}}{S_T^2} \qquad [4.53]$$

Siendo:

S_j^2 = varianza del subtest j.

r_{jj} = coeficiente de fiabilidad del subtest j.

S_T^2 = varianza de la batería total.

12. EJERCICIOS DE AUTOEVALUACIÓN

1. La razón entre la desviación típica de los errores y la desviación típica de las puntuaciones empíricas es 0,45. ¿Cuál es el valor del coeficiente de fiabilidad?
2. Calcular el coeficiente de fiabilidad de un test sabiendo que la varianza de las puntuaciones empíricas es igual a 36 y el error típico de medida es 3.
3. ¿Cuál es el valor del coeficiente de fiabilidad si la proporción de varianza verdadera que hay en la varianza empírica de un test es 0,90?
4. Hemos aplicado un test a un grupo de 100 sujetos. La desviación típica de los errores de medida es 2, lo que significa el 10% de la varianza de las puntuaciones verdaderas. Calcular el coeficiente de fiabilidad de dicho test.
5. Hemos aplicado un test de fluidez verbal a un grupo de 150 sujetos. La varianza de las puntuaciones empíricas de los sujetos de dicho grupo fue 36 y el coeficiente de fiabilidad 0,85.

   Calcular:

   a) El error típico de medida del test.

   b) Utilizando el modelo de regresión el intervalo confidencial dentro del cual podemos afirmar que se encontrará la puntuación diferencial verdadera de un sujeto cuya puntuación típica empírica fue de 0,75, utilizando para ello el modelo de regresión (N.C. 99%).
6. El Instituto Nacional de Calidad desea examinar el nivel de conocimientos en el área de Humanidades de los alumnos al finalizar la educación obligatoria. Para ello, construye una prueba de cinco preguntas cortas, calificadas en una escala de 1 a 5 cada una de ellas; esta prueba se administra a una muestra representativa de 2.000 alumnos procedentes de todas las comunidades autónomas. En la tabla adjunta se presentan las respuestas dadas a las preguntas de dicha prueba por los seis primeros alumnos de la muestra.

   Calcular:

   a) La fiabilidad de la prueba.

   b) Si se añadieran a la prueba 5 preguntas paralelas a las ya existentes, ¿se obtendría un coeficiente de fiabilidad significativamente diferente al anterior? La correlación entre las puntuaciones del test original y del alargado es 0,85 (N.C. 95%).

   c) Estimar la puntuación verdadera en el test original del alumno número 4.

ALUMNOS	PREGUNTAS				
	1	2	3	4	5
1	3	2	4	3	4
2	2	3	4	3	2
3	5	4	3	4	5
4	2	1	2	2	1
5	3	2	2	1	3
6	4	5	4	5	4

7. Ejercicios conceptuales

   A continuación se ofrecen una serie de enunciados ante los que tendrá que responder si son verdaderos o falsos:

   1. Si dos tests son paralelos, las medias de las puntuaciones empíricas deben ser iguales.
   2. El coeficiente de fiabilidad expresa la proporción de la varianza verdadera que hay en la varianza de las puntuaciones empíricas.
   3. El coeficiente α es un índice de la estabilidad de las medidas.
   4. Un test tiene un único coeficiente de fiabilidad.
   5. En el caso de que un test esté formado por ítems dicotómicos de igual nivel de dificultad, el mejor estimador del coeficiente de fiabilidad lo constituye la ecuación KR21.
   6. Si un test tiene un coeficiente de fiabilidad igual 0,80, el índice de fiabilidad es igual a 0,64.
   7. Si se cumple que $S_v^2 = S_x^2$ el coeficiente de fiabilidad $r_{xx} = 1$.
   8. Para calcular la fiabilidad de un test mediante el método de dos mitades, aplicamos el test una sola vez.
   9. En la fórmula de Spearman-Brown, *n* indica el número de ítems del test.
   10. Se define el error típico de medida como la desviación típica de los errores de medida.
   11. El coeficiente de fiabilidad de un test es igual a cero si $S_e^2 = 0$.
   12. El coeficiente de fiabilidad varía entre –1 y 1.
   13. La fiabilidad de un test depende de la longitud del mismo.
   14. El valor de $\alpha \leq r_{xv}^2$.
   15. La correlación entre las puntuaciones empíricas y los errores es siempre igual cero.

13. SOLUCIONES A LOS EJERCICIOS DE AUTOEVALUACIÓN

1.

$$\frac{S_e}{S_x} = 0,45$$

$$r_{xx} = 1 - \frac{S_e^2}{S_x^2} = 1\text{-}0,45^2 = 1 - 0,20 = 0,80$$

2.

$$S_e = S_x\sqrt{1-r_{xx}}; \quad S_e^2 = S_x^2 - S_x^2 r_{xx}; \quad 36r_{xx} = 36 - 9; \quad r_{xx} = \frac{27}{36} = 0,75$$

o también

$$r_{xx} = 1 - \frac{S_e^2}{S_x^2} = 1 - \frac{9}{36} = 1 - 0,25 = 0,75$$

3.

$$r_{xx} = r_{xv}^2 = \frac{S_v^2}{S_x^2} = 0,90$$

4. $S_e = 2$

$$\frac{S_e}{S_v^2} = 0,10; \quad S_v^2 = \frac{2}{0,10} = 20$$

$$S_x^2 = S_v^2 + S_e^2 = 20 + 4 = 24; \quad r_{xx} = \frac{S_v^2}{S_x^2} = \frac{20}{24} = 0,83$$

5.

a) $S_e = S_x\sqrt{1-r_{xx}} = 6\sqrt{1-0,85} = 2,32$

b) N.C. 99% $\rightarrow Z_c = \pm 2,58$

$Z = 0,75 \rightarrow x = 0,75 \cdot 6 = 4,5$

$S_{vx} = S_e\sqrt{r_{xx}} = 2,32\sqrt{0,85} = 2,14$

$v' = r_{xx} \cdot x = 0,85 \cdot 4,5 = 3,82$

$E_{máx} = 2,58 \cdot 2,14 = 5,52$

$3,82 \pm 5,52 \rightarrow -1,70 \leq v \leq 9,34$

6.

	Preguntas											
Alumnos	X_1	X_2	X_3	X_4	X_5	X	X_1^2	X_2^2	X_3^2	X_4^2	X_5^2	X^2
1	3	2	4	3	4	16	9	4	16	9	16	256
2	2	3	4	3	2	14	4	9	16	9	4	196
3	5	4	3	4	5	21	25	16	9	16	25	441
4	2	1	2	2	1	8	4	1	4	4	1	64
5	3	2	2	1	3	11	9	4	4	1	9	121
6	4	5	4	5	4	22	16	25	16	25	16	484
	19	17	19	18	19	92	67	59	65	64	71	1562

a)

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} S_i^2}{S_X^2}\right)$$

$S_1^2 = 67/6 - (19/6)^2 = 1,14$

$S_2^2 = 59/6 - (17/6)^2 = 1,81$

$S_3^2 = 65/6 - (19/6)^2 = 0,81$

$S_4^2 = 64/6 - (18/6)^2 = 1,67$

$S_5^2 = 71/6 - (19/6)^2 = 1,81$

$S_X^2 = 1562/6 - (92/6)^2 = 25,22$

$$\alpha = \frac{5}{5-1}\left(1-\frac{1,14+1,81+0,81+1,67+1,81}{25,\ 22}\right) = 0,89$$

Teniendo en cuenta el resultado obtenido, podemos concluir que el test constituye un buen instrumento para medir el nivel de conocimientos en el área de Humanidades.

b)

$$n = EF / EI = 10/5 = 2$$

$$r_{xx} = \frac{2 \cdot 0,89}{1+0,89} = 0,94$$

$$t = \frac{|\hat{\alpha}_1 - \hat{\alpha}_2|\sqrt{N-2}}{\sqrt{4(1-\hat{\alpha}_1)(1-\hat{\alpha}_2)(1-r_{12}^2)}} \equiv T_{N-2}$$

$$t = \frac{|0,94-0,89|\sqrt{6-2}}{\sqrt{4(1-0,94)(1-0,89)(1-0,85^2)}} = 1,17 < t_{95,4} = 2,78$$

No parecen existir diferencias estadísticamente significativas entre las pruebas de 5 y 10 preguntas, a ese nivel de confianza.

c)

$$V' = r_{xx}(X - \overline{X}) + \overline{X}$$

$$V' = 0,89(8-15,33) + 15,33 = 8,81$$

7. Soluciones a los ejercicios conceptuales

1. El enunciado es verdadero.

   Teniendo en cuenta que la esperanza matemática de los errores de medida es cero y que las puntuaciones verdaderas de los sujetos son iguales en ambos tests, podemos concluir la existencia de igualdad entre las medias de las puntuaciones empíricas.

2. El enunciado es verdadero.

   Se expresa como el cociente entre la varianza de las puntuaciones verdaderas y la varianza de las puntuaciones empíricas y se puede interpretar como la proporción de la varianza de las puntuaciones empíricas de los sujetos que se debe a la varianza verdadera o lo que es lo mismo, la proporción de varianza verdadera que hay en la varianza empírica.

3. El enunciado es falso.

   El coeficiente α es un estimador de la consistencia interna del test.

4. El enunciado es falso.

   El valor del coeficiente de fiabilidad no depende únicamente de las características propias del test, sino de otros factores como la variabilidad de la muestra en la que es aplicado y la longitud del test.

5. El enunciado es verdadero.

6. El enunciado es falso.

   $$r_{xv} = \sqrt{r_{xx}} = \sqrt{0,80} = 0,89$$

7. El enunciado es verdadero.

   $$S_x^2 = S_v^2 + S_e^2 \quad \text{si} \quad S_x^2 = S_v^2 \Rightarrow S_e^2 = 0 \qquad r_{xx} = \frac{S_v^2}{S_x^2} = 1$$

8. El enunciado es verdadero.

9. El enunciado es falso.

   «*n*» indica el número de veces que hay que alargar o reducir la longitud del test.

10. El enunciado es verdadero.

11. El enunciado es falso.

    $r_{xx} = 1$, puesto que $r_{xx} = 1 - \frac{S_e^2}{S_x^2}$

12. El enunciado es falso.

    El coeficiente de fiabilidad varía entre 0 y 1. Definimos el coeficiente de fiabilidad como el cociente entre la varianza de las puntuaciones verdaderas y la varianza de las puntuaciones empíricas. Esta forma de expresar el coeficiente de fiabilidad nos indica la proporción de la varianza verdadera que se puede explicar a partir de la varianza empírica de las puntuaciones de los sujetos. A medida que dicha proporción aumenta, disminuye el error de medida. Si $r_{xx'} = 1$, el error es cero lo que implica una fiabilidad perfecta del test. Sin embargo, a medida que dicha proporción disminuye se produce un incremento en el error de medida. En el caso de que $r_{xx'} = 0$, la varianza de los errores de medida sería igual a la varianza de las puntuaciones empíricas.

13. El enunciado es verdadero.

    Uno de los factores que influye en la fiabilidad de un test es su longitud, es decir, el número de ítems que lo componen. Cuantos más ítems representativos del rasgo a medir se utilicen mayor será la información que obtengamos acerca del atributo que estemos estudiando y, consiguientemente, cabe pensar que menor será el error que cometamos al pronosticar la puntuación verdadera de un sujeto. Por lo tanto, la fiabilidad del test se incrementará. Ahora bien, llega un momento en que por más que se aumente el número de ítems ya no se produce un aumento significativo.

14. El enunciado es verdadero.

    El coeficiente alpha puede ser considerado como una estimación del límite inferior del coeficiente de fiabilidad de un test.

15. El enunciado es falso.

    Esta correlación viene expresada como: $r_{xe} = \sqrt{1-r_{xx}}$. La correlación así establecida es igual o mayor de cero. Su valor máximo se alcanzará cuando la fiabilidad del test es nula ($r_{xx} = 0$) y su valor mínimo se alcanzará cuando la fiabilidad del test es perfecta ($r_{xx} = 1$).

14. APÉNDICE

A continuación se ofrecen las demostraciones de las fórmulas que han ido apareciendo a lo largo del tema.

4.3

$r_{ve} = \frac{\sum ve}{NS_vS_e} = \frac{\sum ve}{N}\frac{1}{S_vS_e}$. Como $\frac{\sum ve}{N} = 0 \Rightarrow r_{ve} = 0$

4.5

$E = X - V$

Por definición, la ecuación del modelo establece que: $X = V + E$. Despejando: $E = X - V$

4.6

$E(e) = 0$

$e = X - V$, luego la $E(e) = E(X) - E(V)$. Según el primer supuesto del modelo sabemos que: $E(X) = V$, por lo tanto: $E(e) = V - E(V) = V - V = 0$

4.8

Dado que la covarianza es, $Cov\ (v,e) = r_{ve}\ S_x\ S_e$, y, según el segundo supuesto, $r_{ve} = 0$ podemos inferir que $Cov\ (v,e) = 0$

4.9

La varianza de una variable que es suma de otras dos es igual a la suma de las varianzas de cada una de las variables más el doble de las covarianzas. $S_x^2 = S_{(v+e)}^2 = S_v^2 + S_e^2 + 2Cov(v,e)$.

Partiendo del segundo supuesto del modelo sabemos que, $r_{ve} = \frac{Cov(v,e)}{S_xS_e}$, de donde podemos concluir que el valor de $Cov(v,e) = 0$. Por lo tanto $S_x^2 = S_v^2 + S_e^2$

4.10

$Cov(X,V) = S_v^2$

La $Cov(X,V) = E(XV) - E(X)E(V)$. Según el modelo lineal $X = V + e$, sustituyendo

$Cov(X,V) = E((V+e)V) - E(V+e)E(V) = E(V)^2 + E(Ve) - E(V)E(V) - E(e)E(V)$

Puesto que: $E(Ve)-E(e)E(V) = Cov(v,e)$, y la $Cov(V,e) = 0$, podemos establecer,

$Cov(X,V)=E(V^2)-(E(V))^2 = S_v^2$

4.11

$$r_{xe} = \frac{S_e}{S_x}$$

En puntuaciones diferenciales:

$$r_{xe} = \frac{\sum xe}{NS_xS_e} = \frac{\sum (v+e)e}{NS_xS_e} = \frac{\sum ve + \sum e^2}{NS_xS_e} = \frac{\sum ve}{NS_xS_e} + \frac{\sum e^2}{NS_xS_e}$$

como $\frac{\sum ve}{N} = r_{ve}S_vS_e$, y $\frac{\sum e^2}{N} = S_e^2$, podemos establecer que

$r_{xe} = \frac{r_{ve}S_vS_e}{S_xS_e} + \frac{S_e^2}{S_xS_e} = \frac{S_e}{S_x}$ ya que $r_{ve}\,S_v\,S_e = 0$, por ser igual a la covarianza entre las puntuaciones verdaderas y los errores.

4.12

$Cov\,(X_1,X_2) = Cov\,(V_1,V_2) = E(X_1,X_2) - E(X_1)\,E(X_2)$. Según el modelo lineal $X = V + e$, sustituyendo en X_1 y X_2.

$Cov\,(X_1,X_2) = E((V_1+e_1)\,(V_2+e_2)) - E(V_1+e_1)\,E(V_2+e_2) = E(V_1V_2) + E(V_1e_2) + E(e_1V_2) + E(e_1e_2) - - E(V_1)\,E(V_2) - E(V_1)\,E(e_2) - E(e_1)\,E(V_2) - E(e_1)\,E(e_2)$

Como: $E(V_1e_2) - E(V_1)\,E(e_2) - Cov\,(V_1,e_2) = 0$

$E(e_1V_2) - E(e_1)\,E(V_2) - Cov\,(e_1,V_2) = 0$

$E(e_1e_2) - E(e_1)\,E(e_2) - Cov\,(e_1,e_2) = 0$

Es decir, no existe covariación entre las puntuaciones verdaderas y los errores, y tampoco entre los errores entre sí, por lo que podemos concluir que:

$Cov\,(X_1,X_2) = E(V_1V_2) - E(V_1)\,E(V_2) = Cov\,(V_1,V_2)$

Si tenemos formas paralelas entonces, $Cov\,(X_1,X_2) = Cov\,(V_1,V_2) = \text{Var}\,(V)$

4.13

Por definición sabemos que la correlación entre las puntuaciones obtenidas por una muestra de sujetos en dos formas paralelas la podemos expresar como $r_{xx'} = \frac{Cov(X,X')}{S_xS_{x'}}$. Según la expresión (3.10): $Cov\,(X,X') = S_v^2$. Asímismo, hemos establecido que las varianzas de las puntuaciones empíricas en dos tests paralelos son iguales, luego podemos establecer la igualdad: $S_X = S_{X'}$ y que $S_X\,S_X = S_x^2$. De donde se concluye que $r_{xx'} = \frac{S_V^2}{S_X^2} = r_{xV}^2$.

4.14

Como consecuencia de la expresión 4.13, se deduce fácilmente que
$r_{x_1x_2} = r_{x_1x_3} = r_{x_2x_3} = \ldots\ldots = r_{x_jx_k}$

Sabemos que la correlación entre dos formas paralelas de un test (X, X') puede expresarse como: $r_{xx'} = \frac{Cov(X,X')}{S_xS_{x'}}$. Según hemos visto $Cov\,(X,X') = S_v^2$ y, por ser formas paralelas, $S_x = S_{x'}$

Podemos establecer $r_{xx'} = \frac{S_V^2}{S_X^2}$ y que el coeficiente de fiabilidad, dados dos o mas tests paralelos, es el mismo para todos puesto que se mantiene constante tanto el valor de la varianza verdadera como el de la varianza empírica.

4.17

$$r_{xx'} = 1 - r_{xe}^2$$

$r_{xx'} = \frac{S_v^2}{S_x^2}$, puesto que $S_x^2 = S_v^2 + S_e^2$ tenemos, $\frac{S_x^2 - S_e^2}{S_x^2} = 1 - \frac{S_e^2}{S_x^2} = 1 - r_{xe}^2$

4.18

$$r_{xe} = \sqrt{1 - r_{xx'}}$$

$$r_{xe} = \frac{S_e}{S_x} = \sqrt{\frac{S_e^2}{S_x^2}} = \sqrt{\frac{S_x^2 - S_v^2}{S_x^2}} = \sqrt{1 - \frac{S_v^2}{S_x^2}} = \sqrt{1 - r_{xx'}}$$

4.19

$$S_e = S_x\sqrt{1-r_{xx'}}$$

Según hemos visto, $r_{xx'} = 1-\frac{S_e^2}{S_x^2}$, operando $r_{xx'} = \frac{S_x^2 - S_e^2}{S_x^2} \Rightarrow r_{xx'}S_x^2 = S_x^2 - S_e^2$,

despejando S_e^2 tenemos $S_e^2 = S_x^2 - S_x^2 r_{xx'} = S_x^2 - S_x^2(1-r_{xx'})$ de donde: $S_e = S_x\sqrt{1-r_{xx'}}$

4.20

$$S_{vx} = S_x\sqrt{1-r_{xx'}}\sqrt{r_{xx'}} = S_e\sqrt{r_{xx'}}$$

En puntuaciones diferenciales podemos expresar: $S_{vx}^2 = \frac{\sum(v-v')^2}{n}$

Mediante la ecuación de regresión en puntuaciones diferenciales $v' = r_{xx}x$,

Sustituyendo:

$$S_{vx}^2 = \frac{\sum(v-v')^2}{n} = \frac{\sum(v-r_{xx}x)^2}{n} = \frac{\sum(v^2+(r_{xx}x)^2-2r_{xx}xv)}{n} = \frac{\sum v^2}{n} + r_{xx}^2\frac{\sum x^2}{n} - 2r_{xx}\frac{\sum xv}{n} =$$
$$= S_v^2 + r_{xx}^2S_x^2 - 2r_{xx}S_vS_vr_{xv}$$

Como hemos visto: $r_{xv} = \frac{S_v}{S_x}$ y $r_{xx} = \frac{S_v^2}{S_x^2}$ de donde:

$S_v^2 + r_{xx}S_v^2 - 2r_{xx}S_vS_x\frac{S_v}{S_x} = S_v^2 + r_{xx}S_v^2 - 2r_{xx}S_v^2 = S_v^2 - S_v^2r_{xx} = S_v^2(1-r_{xx})$, teniendo en cuenta que,

$S_v = S_x\sqrt{r_{xx}}$ sustituyendo en la expresión anterior:

$$S_{vx} = S_x\sqrt{1-r_{xx'}}\sqrt{r_{xx'}} = S_e\sqrt{r_{xx'}}$$

4.21

$$S_{X_1-X_2} = S_x\sqrt{1-r_{xx'}}\sqrt{2}$$

Por lo general, la puntuación V' estimada a partir de las ecuaciones de la recta de regresión no coincide con la puntuación verdadera del sujeto V. La diferencia entre la puntuación verdadera del sujeto (V) y la puntuación verdadera estimada (V') es lo que conocemos como el

error de estimación. Definimos el error típico de estimación (S_{vx}), como la desviación típica de los errores de estimación.

$$S_{x_1-x_2}^2 = S_{e_x}^2 = \frac{\sum(x_1-x_2)^2}{N} = \frac{\sum x_1^2}{N} + \frac{\sum x_2^2}{N} - 2\frac{\sum x_1x_2}{N} = S_{x_1}^2 + S_{x_2}^2 - 2r_{x_1x_2}S_{x_1}S_{x_2}$$

Según el modelo, las varianzas en los tests paralelos son iguales por lo que:

$S_{x_1-x_2} = 2S_x^2 - 2r_{xx'}S_x^2 = 2S_x^2(1-r_{xx'})$ simplificando y sacando la raíz cuadrada:

$$S_{x_1-x_2} = S_x\sqrt{1-r_{xx'}}\sqrt{2}$$

4.24

$$R_{xx} = \frac{nr_{xx}}{1+(n-1)r_{xx}}$$

Partimos de la definición del coeficiente de fiabilidad como cociente entre la varianza verdadera y la varianza empírica de las puntuaciones de los sujetos en un test.

$$R_{xx} = \frac{S_{nv}^2}{S_{nx}^2}$$

A continuación descomponemos tanto la varianza verdadera como la varianza empírica del test total.

La varianza de las puntuaciones verdaderas, S_{nv}^2, será igual a la suma de las «n» varianzas de las puntuaciones verdaderas más la suma de las «$n(n-1)$» covarianzas: $S_{nv}^2 = \sum S_{v_a}^2 + \sum r_{v_av_b}S_{v_a}S_{v_b}$. Puesto que partimos del supuesto de ítems paralelos, tanto las varianzas como las covarianzas son iguales, por lo que la expresión anterior puede formularse como: $S_{nv}^2 = nS_{v_a}^2 + n(n-1)r_{v_av_b}S_{v_a}S_{v_b}$. También sabemos que la correlación $r_{vavb} = 1$, ya que es la correlación entre las puntuaciones verdaderas, y que $S_{va} = S_{vb}$, por lo tanto: $S_{nv}^2 = nS_{v_a}^2 + n(n-1)S_{v_a}^2$.

Sacando factor común a $nS_{v_a}^2$, $S_{nv}^2 = nS_{v_a}^2(1+(n-1))$. Simplificando podemos concluir que la varianza de las puntuaciones verdaderas en el test total puede expresarse: $S_{nv}^2 = nS_{na}^2$

Veamos ahora lo que ocurre en el caso de la varianza empírica, S_{nx}^2. La varianza de las puntuaciones empíricas será igual a la suma de las «n» varianzas de las puntuaciones empíricas más la suma de las «$n(n-1)$» covarianzas: $S_x^2 = \sum S_{x_a}^2 + \sum r_{x_ax_b}S_{x_a}S_{x_b}$. Puesto que partimos del

supuesto de ítems paralelos, tanto las varianzas como las covarianzas son iguales entre, por lo que la expresión anterior puede formularse como $S^2_{nx} = nS^2_{x_a} + n(n-1)r_{x_a x_b}S^2_{x_a}$, por ser $S_{x_a} = S_{x_b}$

Sacando factor común a $nS^2_{x_a}$, $S^2_{nx} = nS^2_{x_a}(1+(n-1)r_{x_a x_b})$.

Sustituyendo el valor de la varianza verdadera y la varianza empírica en la expresión del coeficiente de fiabilidad, tenemos:

$R_{xx} = \frac{S^2_{nv}}{S^2_{nx}} = \frac{nS^2_{v_a}}{nS^2_{x_a}(1+(n-1)r_{x_a x_b})} = n\frac{S^2_{v_a}}{S^2_{x_a}}\frac{1}{1+(n-1)r_{x_a x_b}}$. Si tenemos en cuenta que $\frac{S^2_{v_a}}{S^2_{x_a}} = r_{x_a x_a}$ y que las intercorrelaciones entre cada dos o más tests paralelos son iguales, es decir, $r_{xaxa} = r_{xaxb} = r_{xx}$, podemos concluir: $R_{xx} = \frac{nr_{xx}}{1+(n-1)r_{xx}}$

Partiendo de lo anterior, cuando se aumenta «n» veces la longitud del *test la varianza de los errores* sería: $S^2_{ne} = nS^2_e + n(n-1)r_{xx}$

Un caso particular de esta fórmula es cuando se duplica la longitud del test inicial.

En ocasiones lo que pretendemos es que un test tenga una determinada fiabilidad, y lo que nos planteamos es saber cuántos ítems tendríamos que aumentar el test para conseguir dicho coeficiente.

El número de ítems que tenemos que aumentar dicho test lo podemos hallar despejando el término «n» de la ecuación general de Spearman-Brown.

$$R_{xx} = \frac{nr_{xx}}{1+(n-1)r_{xx}} = \frac{nr_{xx}}{1+nr_{xx}-r_{xx}}$$

$$R_{xx} + nr_{xx}R_{xx} - r_{xx}R_{xx} = nr_{xx}; \quad R_{xx} - r_{xx}R_{xx} = nr_{xx} - nr_{xx}R_{xx} = n(r_{xx} - r_{xx}R_{xx})$$

$$n = \frac{R_{xx} - r_{xx}R_{xx}}{r_{xx} - r_{xx}R_{xx}} = \frac{R_{xx}(1-r_{xx})}{r_{xx}(1-R_{xx})}$$

Una vez conocido el valor de «n» podemos calcular el número de elementos finales (EF). $EF = EI \cdot n$. La diferencia entre los ítems finales y los ítems iniciales nos dará el número de elementos que habría que añadir o disminuir un test para obtener el coeficiente de fiabilidad deseado.

4.28

$$R_{xx} = \frac{2r_{xx}}{1+r_{xx}}$$

Supongamos que tenemos una serie de formas paralelas y que juntamos éstas de dos en dos: $x_a + x_b$, $x_c + x_d$.

Puesto que dichos tests son paralelos podemos establecer: $r_{xx} = r_{ab} = r_{ac} = \ldots = r_{cd}$, es decir, dados dos o mas tests paralelos, las intercorrelaciones entre cada dos de ellos son iguales.

Por definición el coeficiente de fiabilidad del test R_{xx} puede expresarse, en puntuaciones diferenciales como: $R_{xx} = \frac{\sum(x_a + x_b)(x_c + x_d)}{NS_{(x_a+x_b)}S_{(x_c+x_d)}}$, al ser formas paralelas, las desviaciones típicas serán iguales ($S_{(x_a+x_b)} = S_{(x_c+x_d)}$), por lo que podemos expresar el denominador como $S^2_{(x_a+x_b)}$, y sustituyendo: $R_{xx} = \frac{\sum(x_a + x_b)(x_c + x_d)}{N}\frac{1}{S^2_{(x_a+x_b)}}$.

Si desarrollamos el primer término tendremos:

$$\frac{\sum(x_a + x_b)(x_c + x_d)}{N} = \frac{\sum x_a x_c}{N} + \frac{\sum x_a x_d}{N} + \frac{\sum x_b x_c}{N} + \frac{\sum x_b x_d}{N}$$

puesto que estos cuatro términos expresan covarianza, les podemos sustituir por $r_{ac}S_a S_c + r_{ad}S_a S_d + r_{bc}S_b S_c + r_{bd}S_d S_d$ y, al ser formas paralelas, la expresión puede escribirse como: $4S^2_x r_{xx}$.

Si desarrollamos el término $S^2_{(x_a+x_b)}$, puesto que la varianza de una variable que es suma de otras dos es igual a la suma de las varianzas de cada una de las variables más el doble de las covarianzas:

$$S^2_{(x_a+x_b)} = \frac{\sum(x_a + x_b)^2}{N} = \frac{\sum x_a^2}{N} + \frac{\sum x_b^2}{N} + 2\frac{\sum x_a x_b}{N} =$$

$$= S^2_a + S^2_b + 2r_{ab}S_a S_b = 2S^2_x + 2r_{xx}S^2_x = 2S^2_x(1+r_{xx})$$

Sustituyendo,

$$R_{xx} = \frac{4S_x^2 r_{xx}}{2S_x^2(1+r_{xx})}$$

y simplificando,

$$R_{xx} = \frac{2r_{xx}}{1+r_{xx}}$$

Esta misma expresión puede obtenerse a partir de la influencia del aumento de la longitud de un test sobre la varianza verdadera, la varianza empírica y la varianza de error.

En primer lugar veamos como se ve afectada la varianza de las puntuaciones empíricas de los sujetos, cuando se duplica la longitud del test. Supuesto los ítems paralelos, las varianzas de las dos mitades son iguales, es decir, $S_a^2 = S_b^2$, con lo que la varianza total del test puede expresarse como, S_{2x}^2.

Puesto que, como ya hemos dicho, la varianza de una variable que es suma de otras dos es igual a la suma de las varianzas de cada una de las variables más el doble de las covarianzas, tendremos:

$$S_{2x}^2 = S_a^2 + S_b^2 + 2r_{ab}S_aS_b \text{ de donde, } S_{2x}^2 = 2S_x^2(1+r_{xx})$$

Veamos ahora lo que sucede respecto a la varianza verdadera. La varianza de la distribución de las puntuaciones verdaderas, S_v^2, puede expresarse como $S_{2v}^2 = S_{v_a}^2 + S_{v_b}^2 + 2r_{v_av_b}S_{v_a}S_{v_b}$

Las puntuaciones verdaderas en los dos tests paralelos son iguales y la correlación $r_{v_av_b} = 1$, ya que es la correlación entre las puntuaciones verdaderas. Luego:

$$S_{2v}^2 = S_{v_a}^2 + S_{v_b}^2 + 2S_{v_a}S_{v_b} = 4S_v^2 \text{ , puesto que } S_{v_a} = S_{v_b}$$

Es decir, cuando se duplica la longitud de un test dado, la varianza de las puntuaciones verdaderas de los sujetos es igual a cuatro veces la varianza de las puntuaciones verdaderas de cada una de las mitades.

Por último, veamos lo que sucede respecto a la varianza de error. Por ser tests paralelos, partimos del supuesto de que las varianzas $S_{e_a}^2 = S_{e_b}^2$ y que la correlación entre los errores $r_{e_ae_b} = 0$. La varianza de error (S_e^2) puede expresarse en los siguientes términos: $S_{2e}^2 = S_{e_a}^2 + S_{e_b}^2 + 2r_{e_ae_b}S_{e_a}S_{e_b} = 2S_e^2$, puesto que la covarianza se anularía al ser la correlación entre errores igual a cero.

Dado que el coeficiente de fiabilidad (r_{xx}) es igual al cociente entre la varianza verdadera (S_v^2) y la varianza empírica (S_x^2), tendremos que el coeficiente de fiabilidad, al duplicar la longitud del test viene expresado por:

$$R_{xx} = \frac{S_{2v}^2}{S_{2x}^2} = \frac{4S_v^2}{2S_x^2(1+r_{xx})} = \frac{2r_{xx}}{1+r_{xx}}$$

que es precisamente la expresión a la cual habíamos llegado anteriormente.

4.31

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum S_j^2}{S_x^2}\right)$$

Como ya hemos comentado, la varianza de una variable, suma de «*n*» variables, es definida como la suma de las varianzas, más la suma de las covarianzas, con lo que la varianza total de las puntuaciones empíricas de los sujetos en un test la podemos expresar como:

$$S_x^2 = \sum S_j^2 + \sum\sum r_{jk}S_jS_k$$

es decir, suma de las varianzas de cada uno de los elementos más la de las covarianzas entre todos ellos.

Si los ítems son paralelos, se puede calcular la varianza media y la covarianza media de los items.

$$\sum S_j^2 = n\overline{S_j^2} \text{ , ya que } \overline{S_j^2} = \frac{\sum S_j^2}{n}$$

$$\sum_j^n \sum_k^n r_{jk}S_jS_k = n(n-1)\overline{r_{jk}S_j^2}$$

de donde:

$$S_x^2 = \sum^n S_j^2 + n(n-1)\overline{r_{jk}S_j^2}$$

y despejando:

$$\overline{r_{jk}} = \frac{S_x^2 - \sum^n S_j^2}{(n-1)\sum S_j^2}$$

En el caso de querer estimar la fiabilidad del test total aplicaremos la ecuación general de Spearman-Brown para el caso de un test de longitud «*n*».

$$r_{xx} = \frac{n \cdot \overline{r_{jk}}}{1+(n-1)\overline{r_{jk}}}$$

donde «n» representa el número de ítems y, $\overline{r_{jk}}$ representa la correlación promedio de las $n(n-1)$ correlaciones entre los ítems. Si lo sustituimos por la expresión anterior:

$$r_{xx} = \frac{S_x^2 - \sum^n S_j^2}{n \cdot \frac{(n-1)\sum S_j^2}{1+(n-1)\frac{S_x^2 - \sum S_j^2}{(n-1)\sum S_j^2}}}$$

despejando,

$$\alpha = \frac{n}{n-1}\left(\frac{S_x^2 - \sum S_j^2}{S_x^2}\right) = \frac{n}{n-1}\left(1 - \frac{\sum S_j^2}{S_x^2}\right)$$

4.33

$\bar{\alpha} = \hat{\alpha}$, cuando $n \to \infty$

$$\bar{\alpha} = \frac{(N-3)\hat{\alpha}+2}{N-1} = \frac{(N-3)\hat{\alpha}}{N-1} + \frac{2}{N-1} = \frac{N\hat{\alpha}-3\hat{\alpha}}{N-1} + \frac{2}{N-1} =$$

$$= \frac{N\hat{\alpha}}{N-1} - \frac{3\hat{\alpha}}{N-1} + \frac{2}{N-1} = \frac{\hat{\alpha}}{\frac{N-1}{N}} - \frac{3\hat{\alpha}}{N-1} + \frac{2}{N-1} =$$

$$= \frac{\hat{\alpha}}{(N/N)-(1/N)} - \frac{3\hat{\alpha}}{N-1} + \frac{2}{N-1} = \frac{\hat{\alpha}}{1-(1/n)} - \frac{3\hat{\alpha}}{N-1} + \frac{2}{N-1}$$

Si $n \to \infty$ entonces $1/N = 0$, $3\alpha/N - 1 = 0$, $2/N - 1 = 0$; de donde podemos deducir que $\bar{\alpha} = \hat{\alpha}$

— *Relación entre la ecuación de Rulon y la ecuación de Guttman-Flanagan*

$r_{xx} = 1 - \frac{S_{p-i}^2}{S_x^2} = \frac{S_x^2 - S_{p-i}^2}{S_x^2}$, puesto que la varianza de una variable que es suma de otras dos es igual a la suma de las varianzas de cada una de las variables más el doble de las covarianzas, tenemos:

$$\frac{S_p^2 + S_i^2 + 2r_{pi}S_pS_i - S_p^2 - S_i^2 + 2r_{pi}S_pS_i}{S_x^2} = \frac{4r_{pi}S_pS_i}{S_x^2}$$

Si desarrollamos ahora la ecuación de Guttman, tendremos:

$r_{xx} = 2\left(1 - \frac{S_p^2 + S_i^2}{S_x^2}\right) = 2\left(\frac{S_x^2 - S_p^2 - S_i^2}{S_x^2}\right)$. Puesto que la varianza de una variable, suma de «*n*» variables, es definida como la suma de sus varianzas, más la suma de las covarianzas, podemos establecer que $S_x^2 = S_p^2 + S_i^2 + 2r_{pi}S_pS_i$ de donde

$$r_{xx} = 2\frac{S_p^2 + S_i^2 + 2r_{pi}^1S_pS_i - S_p^2 - S_i^2}{S_x^2} = \frac{4r_{pi}S_pS_i}{S_x^2}$$

Como puede observarse, en ambos casos, llegamos a la misma expresión final.

— *Relación entre «α» y «β»*

$\beta = \alpha$

Bajo este supuesto podemos establecer que: $n = kn_j$, donde:

n = número de subtests

$$1 - \sum_{j=1}^{n}\left(\frac{n_j}{n}\right)^2 = 1 - \sum_{j=1}^{n}\left(\frac{n_j}{k \cdot n_j}\right)^2 = 1 - k\frac{1}{k^2} = \frac{k-1}{k}$$

Sustituyendo en β,

$$\beta = \frac{S_x^2 - \sum_{j=1}^{n} S_j^2}{S_x^2 \frac{k-1}{k}} = \frac{k}{k-1}\frac{S_x^2 - \sum_{j=1}^{n} S_j^2}{S_x^2} = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{n} S_j^2}{S_x^2}\right) = \alpha$$

15. BIBLIOGRAFÍA COMPLEMENTARIA

Martínez-Arias, R.; Hernández Lloreda, Mª J.; Hernández Lloreda, Mª V. (2006). *Psicometría*. Madrid: Alianza editorial.

Martínez-Arias, R.(1995). *Psicometría: Teoría de los Tests Psicológicos y Educativos*. Madrid: Editorial Síntesis.

Muñiz, J. (1998, 2002). *Teoría Clásica de los Tests*. Madrid: Editorial Pirámide.

Muñiz, J. (1996). *Psicometría*. Madrid: Universitas.

Santisteban, C.(1990). *Psicometría. Teoría y práctica en la construcción de tests*. Madrid: Editorial Norma.

TEMA 5

LA FIABILIDAD DE LOS TESTS REFERIDOS AL CRITERIO

Enrique Vila Abad

SUMARIO

1. ORIENTACIONES DIDÁCTICAS

A lo largo de los temas precedentes se ha visto cómo llevar a cabo la construcción de los instrumentos de medición psicológica y, desde el marco de la teoría clásica de los tests, se han planteado distintos procedimientos para evaluar la fiabilidad de las puntuaciones obtenidas al aplicarlos y estimar la puntuación verdadera de los sujetos en la característica medida. Los tests construidos y evaluados con los procedimientos descritos se denominan: *tests referidos a la norma* debido a que el rendimiento de los sujetos se evalúa en referencia a otros sujetos que forman el grupo normativo. Este enfoque de los tests referidos a normas no proporciona, en ocasiones, una información adecuada de la habilidad real de un sujeto sino de su posición relativa respecto a otros sujetos. Supongamos, a modo de ejemplo, que un sujeto puntúa por encima del 80% de sus compañeros en un determinado test. Si deseamos saber la posición relativa de dicho sujeto respecto al rasgo evaluado tendremos que tener información acerca del grado de representatividad de esa muestra. Si estamos hablando de que un sujeto se encuentra en un percentil 80 respecto a una prueba de resolución de problemas, nos podemos plantear cuestiones como, qué tipo de problemas es capaz de resolver, qué tipo de resolución requieren dichos problemas, cuál es el límite de capacidad de resolución de problemas de dicho sujeto, etc. Este tipo de cuestiones puede ser abordado cuando la evaluación de un sujeto no se realiza en función de un grupo normativo, sino cuando tiene lugar en función del número de objetivos logrados por dicho sujeto en dicho test. Hablaremos en este caso de los *tests referidos al criterio*.

En el presente tema pretendemos desarrollar, lo más ampliamente posible, y siempre dentro de las pretensiones del libro, cuatro aproximaciones básicas a la estimación de la fiabilidad de los tests referidos al criterio. Los modelos que aquí presentamos son adecuados para aquellas situaciones en las que la decisión de clasificar a un sujeto dentro o no de un grupo de maestría esté en función de si ha alcanzado o no una determinada puntuación en el test denominada puntuación de corte.

2. DEFINICIÓN Y OBJETIVOS DE LOS TESTS REFERIDOS AL CRITERIO

Los Tests Referidos al Criterio (TRC) tienen sus orígenes en los trabajos de Flanagan (1951) y Nedelsky (1954) que introdujeron el concepto de *estándar absoluto y relativo* respecto a las puntuaciones obtenidas en los tests. La denominación de *Test Referido al Criterio* se debe a Ebel (1962) y su diferenciación respecto a los tests normativos fue establecida por Glaser en 1963. Según Hambleton (1994), las principales causas que generan su aparición son: la necesidad de conocer la eficacia de los programas educativos, el interés por evaluar el nivel de habilidades básicas alcanzado por los sujetos y el clima contrario al uso de los tests que caracterizaba la situación de la sociedad americana en la década de los años sesenta. Durante esta década, se produce una escasez de investigaciones en este campo. Merece destacar, sin embargo, el artículo de Popham y Husek (1969) en el que se reaviva el tema y se amplían las distinciones entre tests referidos al criterio y los tests referidos a normas.

Posteriormente, en la década de los setenta, se incrementó notablemente el número de artículos, monografías, libros y revistas en los que se introducen nuevos términos y modalidades de tests (Berk, 1980; Gray, 1978; Hambleton y col., 1978; Huynh, 1976; Popham, 1978; etc.). Estos autores muestran unanimidad al considerar un test referido al criterio como aquél que intenta establecer el estatus de un sujeto respecto al dominio definido. Destaca el trabajo de Millman (1974) en el que realiza la primera recopilación e integración de los avances en esta temática.

A partir de entonces aparecen sucesivamente manuales especializados elaborados por Bejar (1983), Berk (1980, 1984), Osterlind (1998), Popham (1978) y Roid y Haladyna (1982) entre otros, así como diversos números monográficos en las revistas *Journal of Educational Measurement* (1978, Vol. 15, Nº.4) o *Applied Psychological Measurement* (1980, Vol. 4, Nº.4).

Hacia la segunda mitad de los años 80, se produjo una disminución significativa en la producción de publicaciones dedicada a este tema. Hambleton, (1994) señala que ello fue debido a la irrupción en el contexto educativo del nuevo enfoque denominado *medición auténtica (autenthic measurement)* o *evaluación de la ejecución (performance assessment)*. No obstante, él mismo considera que los términos *medición auténtica* y *evaluación de la ejecución* son simplemente términos alternativos al de *medición referida a criterio*. A finales del siglo XX ya es un tema de gran relevancia en el terreno de la medición psicológica y educativa y prueba de ello son los números monográficos publicados en los últimos años en las revistas *Educational Measurement: Issues and Practice* (1994, Vol. 13, Nº 4) y *Applied Measurement in Education* (1995, Vol. 8, Nº.1 y 1997, Vol. 10, Nº. 1).

Se han propuesto numerosas definiciones para hacer referencia a este tipo de tests, aunque según Hambleton (1988) la más aceptada es la propuesta por Popham (1978):

> Un test referido al criterio se utiliza para evaluar el status absoluto del sujeto con respecto a algún dominio de conductas bien definido.

Teniendo en cuenta esta definición, los TRC no constituyen un nuevo marco teórico en la Teoría de los Tests sino un nuevo enfoque que responde a preguntas y necesidades distintas de los tests referidos a las normas (TRN). En los TRN la finalidad es describir al sujeto en el continuo de algún rasgo, haciendo hincapié en las diferencias individuales y expresando su posición relativa respecto al grupo de sujetos denominado grupo normativo. Desde la perspectiva de los TRC el objetivo es construir y evaluar tests que permitan interpretar las puntuaciones en sentido absoluto, sin referencia a ningún grupo, y describir con mayor precisión los conocimientos, habilidades y destrezas de los sujetos en un dominio concreto de contenidos.

3. DIFERENCIAS ENTRE LOS TESTS REFERIDOS A LA NORMA Y LOS TESTS REFERIDOS AL CRITERIO

En cuanto a la construcción del test, en los TRC se delimita claramente el dominio de contenidos o conductas y el uso pretendido del test, mientras que en los TRN los ítems suelen derivarse de alguna teoría de rasgos y no se hace tanto hincapié en la especificación clara del dominio de contenidos. De este modo, en los TRC se presta mucha atención a las especificaciones de contenido y a la elaboración y análisis cualitativo de los ítems. Una descripción detallada del proceso de construcción de un TRC aparece en los trabajos de Hambleton y Rogers (1991) y Popham (1978, 1984) y sobre elaboración de ítems en los de Haladyna (1999), Millman (1984), Osterlind (1998), Popham (1978), Roid y Haladyna (1982) o Shrock y Coscarelli (1989). Por otra parte, la validez de contenido, tal y como se verá en el tema siguiente, es fundamental en este tipo de tests ya que su esencia es la *relevancia* y *representatividad* de los ítems respecto al dominio específico.

También se encuentran diferencias *en los criterios de selección de ítems para el test*. En los TRN el objetivo es maximizar las diferencias individuales por lo que se eligen ítems de dificultad media y alto índice de discriminación para incrementar el poder discriminativo del test. En los TRC, sin embargo, la selección de los ítems se basa en los objetivos del test y en el propósito y finalidad del mismo (Martínez Arias, 1995). Los TRC se pueden utilizar para dos tipos de objetivos: la estimación de la *puntuación dominio* de los sujetos y el *establecimiento de estándares* mediante puntos de corte (Berk, 1980). Un TRC construido atendiendo al primer objetivo se denomina *test referido al dominio* y se utiliza para describir lo que una persona puede hacer en un área de contenido específico. Por otro lado, cuando un test se construye para establecer estándares mediante los puntos de corte, el test se denomina *test de maestría* y es útil para clasificar a los sujetos en una de las posibles categorías de clasificación excluyentes entre sí como *éxito-fracaso*, *apto-no apto* o *trastorno-no trastorno* (Crocker y Algina, 1986).

Según sea el objetivo que se pretenda, la estimación de la fiabilidad de las puntuaciones se realizará de forma diferente (Traub y Rowley, 1980). En este caso, los métodos de la teoría clásica

para tests normativos no son apropiados porque no permiten describir la precisión de las puntuaciones individuales ni la consistencia de las decisiones tomadas a partir de ellas (Hambleton y Rogers, 1994). Nuevos procedimientos han sido necesarios para alcanzar los objetivos de estos tests.

Por otro lado, el establecimiento de estándares lleva consigo la determinación de los puntos de corte que delimitan los estándares. La ubicación de estos puntos de corte ha motivado numerosas investigaciones dada la gran trascendencia que tienen las decisiones que se toman para los sujetos. Una revisión más completa de los criterios y métodos empleados se puede encontrar en Berk (1986, 1996), Cizek (l996), de Gruijter (1985), Faggan (1994), Livingston y Zieky (1982), Shepard, Glaser, Linn y Bohrnstedt (1993) y en los números monográficos de las revistas *Journal of Educational Measurement* (1978, vol. 15, núm. 4) y *Applied Measurement in Education* (1995, vol. 8, núm. l).

Además de la fiabilidad de las clasificaciones y la adecuada ubicación de los puntos de corte, otro aspecto relevante de los *tests de maestría* es la obtención de evidencias acerca de la validez de las decisiones de la clasificación, tal y como se verá en el tema 7. El estudio de este tipo de evidencias se lleva a cabo mediante el análisis de la correspondencia entre las clasificaciones realizadas por el test y las de un criterio de clasificación externo alternativo. Para ello se realiza un proceso de validación referida a un criterio en el que se calcula el coeficiente de validez a través de índices de acuerdo, y se determinan los índices de sensibilidad y especificidad que complementan la información sobre la validez de las decisiones tomadas por el test. Algunos trabajos en esta temática (véase por ejemplo, Dunn, 2000) proponen la aplicación de la Teoría de la Detección de Señales, y más concretamente de las curvas ROC para el estudio de la validez de las decisiones tomadas al clasificar a los sujetos.

Por último, *en lo que respecta a la evaluación de los sujetos*, encontramos diferencias entre ambos enfoques, el normativo y el referido a un criterio, en el significado e interpretación de las puntuaciones de los tests. En los TRN, la puntuación obtenida por los sujetos se considera un indicador de su puntuación verdadera en un rasgo latente y sólo tiene significado en relación a los resultados del grupo normativo. En los TRC, sin embargo, la puntuación representa un estimador del rendimiento del sujeto en el dominio y tiene significado en términos absolutos. En este enfoque, para la estimación de la puntuación en el dominio se puede utilizar la proporción de respuestas correctas (Bock, Thissen y Zimowski, 1997).

4. LONGITUD DEL TEST

El problema de determinar la longitud del test, o el número de ítems que van a evaluar cada uno de los objetivos incluidos en el test, constituye un problema crucial ya que de ello va a depender la utilidad de las puntuaciones obtenidas en dicho test. Si el número de ítems es pequeño, la interpretación que hagamos de las puntuaciones obtenidas tiene un valor limitado. Consiguientemente, se debería ser cauto a la hora de emplear dichas puntuaciones para llevar a cabo cualquier tipo de decisión que implique, por ejemplo, una selección o clasificación de los sujetos. Si tenemos un test con pocos elementos, la estimación del dominio será imprecisa y dará lugar a clasificaciones que o bien son inconsistentes a lo largo de varias presentaciones de formas paralelas, o no son indicativas del verdadero nivel de maestría de un sujeto; es decir, se obtendrán clasificaciones que son poco fiables.

Si el propósito que se persigue es el de poder establecer el grado de maestría de un sujeto, la determinación de la longitud del test está directamente relacionada con el número de errores de clasificación tolerables. Por otra parte, cuando el número de elementos del test es elevado, se pueden asegurar valores de probabilidad de clasificación incorrecta mínimos. Como cabe pensar, un excesivo número de ítems tampoco es lo más adecuado debido a limitaciones de tiempo, economía, etc.

Se pueden considerar dos maneras de reducir el número de errores que se pueden cometer sin tener que aumentar la longitud del test. Por una parte, la utilización de modelos bayesianos (Novick y Jackson, 1974) y, por otra parte, se pueden utilizar métodos basados en tests computarizados (Eignor y Hambleton, 1979; Hambleton y Eignor, 1978; Spineti y Hambleton, 1977; Wilcox, 1980).

A continuación presentamos únicamente el modelo propuesto por Millman (1973). El lector interesado podrá recabar más información a partir de las referencias citadas y/o los trabajos de Birbaum, 1968; Hambleton y col., 1983 y Lord, 1980.

Modelo de Millman

El modelo propuesto por Millman (1973) está basado en el modelo binomial. Considera la proporción esperada de ítems que un sujeto puede contestar correctamente para ser considerado como apto, de la población de ítems definidos, y el error máximo que se está dispuesto a tolerar.

Dicho modelo parte de los siguientes supuestos:

1) El test está compuesto por una muestra aleatoria de ítems dicotómicos.
2) La probabilidad de una respuesta correcta por parte de un sujeto es constante para todos los ítems del test.
3) Las respuestas dadas a los ítems del test son independientes unas de otras.
4) Los errores se ajustan al modelo binomial,

$$\text{Prob}(x \mid p) = \binom{n}{x} p^x q^{n-x} = \sum_{x}^{n} \left(\frac{n!}{x!(n-x)!} \right) p^x q^{n-x} \qquad [5.1]$$

donde:

Prob$(x \mid p)$ = probabilidad de que un sujeto con una puntuación p, conteste correctamente x ítems de un test que tiene n ítems.

A partir de la siguiente ecuación podemos calcular la longitud del test, supuesta una determinada proporción de aciertos:

$$n = \frac{p_c(1-p_c)}{e^2} \qquad [5.2]$$

donde:

n = número de ítems del test.

p_c = proporción de aciertos para ser considerado apto.

e = error máximo admisible.

EJEMPLO:

Para un determinado test se ha establecido la proporción de aciertos para ser considerado apto en 0,85. Se desea saber cuál es la longitud del test si estamos dispuestos a admitir un error máximo de 0,05 y 0,02.

$$n = \frac{0,85(1-0,85)}{0,05^2} = 51 \qquad n = \frac{0,85(1-0,85)}{0,02^2} = 318,75 \approx 319$$

En el primer caso tendríamos 51 ítems y admitiríamos un margen de aciertos entre 0,80 y 0,90 (0,85 ± 0,05) y en el segundo caso tendríamos 319 ítems y un margen de aciertos entre 0,83 y 0,87 (0,85 ± 0,02).

5. FIABILIDAD EN LAS CLASIFICACIONES EN LOS TESTS REFERIDOS AL CRITERIO

Como ya hemos dicho, los tests referidos al criterio se pueden utilizar para dos tipos de objetivos: la estimación de la puntuación dominio de los sujetos, y el establecimiento de estándares mediante puntos de corte (tests de maestría). El segundo enfoque, es el más utilizado y el que ha dado lugar a un mayor número de procedimientos para abordar el problema de la fiabilidad. Es en este contexto desde donde abordaremos el estudio de la fiabilidad de los tests referidos al criterio.

Desde este segundo enfoque, se considera un test fiable si, tras su aplicación a los mismos sujetos en distintas ocasiones, o la aplicación de dos formas paralelas, se clasifica a los sujetos siempre en la misma categoría.

Los métodos que se presentan a continuación para el cálculo de la fiabilidad, se pueden dividir en dos grupos: los que requieren dos aplicaciones del test, y aquellos que sólo requieren una aplicación. Dentro del primer grupo se presenta: el índice de Hambleton y Novick, el coeficiente Kappa de Cohen, y el índice de Crocker y Algina. Dentro del segundo veremos: el método de Huynh, el método de Subkoviak, y el coeficiente de Livingston.

5.1. Índices de acuerdo que requieren dos aplicaciones del test

5.1.1. Coeficiente p_c de Hambleton y Novick

Este coeficiente p_c (Hambleton y Novick, 1973; Swaminathan, Hambleton y Algina, 1974), supone la utilización de la proporción de sujetos que, consistentemente, son clasificados dentro del grupo de maestría o no-maestría, como un índice de la fiabilidad de un test.

Nos basaremos en el siguiente ejemplo para una mayor comprensión de este procedimiento. Supongamos los datos de la tabla 5.1, en la que se presenta la puntuación total obtenida por 20 sujetos en dos tests paralelos compuestos por doce ítems, y que un sujeto debe responder correctamente a un mínimo de 7 ítems para ser clasificado dentro del grupo de maestría.

TABLA 5.1
Puntuación Total

Sujeto	Test «A»	Test «B»	Sujeto	Test «A»	Test «B»
1	7	6	11	5	3
2	9	8	12	5	5
3	8	6	13	4	4
4	8	7	14	3	3
5	7	5	15	4	3
6	6	7	16	3	4
7	6	6	17	2	2
8	6	6	18	5	2
9	6	6	19	3	1
10	5	4	20	1	1

Dichas puntuaciones pueden agruparse tal y como aparecen en la siguiente matriz (Tabla 5.2) en función de que superen o no la puntuación de corte que va a permitir clasificarlos en una categoría u otra.

TABLA 5.2

	Test «B»		
Test «A»	Maestría	No- maestría	Total (N_j)
Maestría	2	3	5
No-maestría	1	14	15
Total (N_i)	3	17	N = 20

Así, los sujetos 2 y 4 son los únicos sujetos que han sido clasificados en el grupo de maestría en ambos tests. Del 7 al 20 los sujetos están clasificados dentro del grupo de no-maestría tanto en el test A como en el B. El resto de los sujetos han sido clasificados de distinta manera en ambos tests. La proporción de sujetos consistentemente clasificados en ambos tests se puede expresar mediante la ecuación:

$$p_c = \sum_{i=1}^{n} p_i = \frac{n_{11}}{N} + \frac{n_{22}}{N} + ... + \frac{n_{mm}}{N} \qquad [5.3]$$

donde:

p_i = proporción de sujetos clasificados consistentemente en ambas formas.

N = número total de sujetos.

n_{11}, n_{22}, ... n_{mm} = número sujetos en cada casilla en los que ambos test coinciden al clasificarlos.

A partir de los datos de la matriz:

$$p_c = \sum_{i=1}^{n} p_i = \frac{2}{20} + \frac{14}{20} = \frac{16}{20} = 0,80$$

El valor máximo de p_c es igual a 1, valor que se obtendrá cuando los sujetos sean clasificados de la misma forma con los dos tests, y el valor mínimo será igual a la proporción de clasificaciones consistentes que podemos esperar por azar (p_a), valor que viene dado en función de las frecuencias marginales de la tabla (N_j).

$$p_a = \sum_{j=1}^{m} \frac{N_j N_i}{N^2} \qquad [5.4]$$

Con los datos de la tabla anterior:

$$p_a = \frac{5 \cdot 3}{20^2} + \frac{15 \cdot 17}{20^2} = 0,0375 + 0,6375 = 0,675 \approx 0,68$$

Ante estos resultados se puede decir que la utilización de los tests supone una mejora importante en la consistencia de las clasificaciones, y por lo tanto en la fiabilidad de las mismas, con respecto a las realizadas por mero azar. Mientras que por azar obtenemos una fiabilidad de 0,68, el uso de los tests nos reporta una fiabilidad de 0,80.

5.1.2. Coeficiente Kappa de Cohen

Esta es una de las medidas más utilizadas cuando se desea estimar el nivel de acuerdo entre varios observadores o jueces. Swaminathan, Hambleton y Algina en 1974 sugieren que en la estimación del coeficiente de fiabilidad se elimine del valor de la proporción de sujetos clasificados consistentemente el valor de la proporción de clasificación consistente esperada por azar y, para ello, recomiendan la utilización del coeficiente Kappa de Cohen (Cohen, 1960; Fleiss y col., 1969; Losada, J. L. y Arnau, J (2000). El coeficiente K de Cohen se puede utilizar cuando clasificamos a los sujetos en dos o más categorías (Muñiz, 1998) cuya fórmula es:

$$K = \frac{p_c - p_a}{1 - p_a} \qquad [5.5]$$

donde:

p_C = proporción de clasificaciones consistentes en ambas formas.

p_a = proporción de clasificaciones consistentes que podemos esperar por azar.

El valor Kappa nos proporciona una medida de la consistencia de clasificación de los sujetos independientemente del posible valor esperado por azar y, tal y como hemos comentado, constituye una de las medidas más utilizadas cuando se desea estimar el nivel de acuerdo entre varios observadores o jueces. El valor del coeficiente Kappa oscila entre -1 y +1. Un valor negativo indicaría situaciones en las cuales existe un total desacuerdo entre los observadores o jueces y, en el contexto de la fiabilidad, carecería de sentido. Un valor de $K = 1$, indicaría una fiabilidad perfecta y un acuerdo perfecto entre los observadores, y un valor de $K = 0$, indicaría que la consistencia observada sería atribuible al azar (Hirji y Rosove, 1990).

No obstante y a pesar de su robustez, este índice deja abierto un interrogante acerca de la significación de los valores obtenidos. Landis y Koch (1977) establecieron la siguiente escala de valoración para el coeficiente kappa como una primera aproximación a la significación de dicho coeficiente.

Coeficiente de kappa	Grado de consistencia
<0.00	Pobre
>0.01 – 0.20	Leve
>0.21 – 0.40	Aceptable
>0.41 – 0.60	Moderado
>0.61 -0.80	Considerable
>0.81 -1.00	Casi perfecto

Si se aplica el coeficiente Kappa a los datos del ejemplo anterior:

$$K = \frac{p_c - p_a}{1 - p_a} = \frac{0,80 - 0,68}{1 - 0,68} = 0,38$$

Con este resultado podríamos hacer una primera interpretación de que el valor obtenido es aceptable.

Este coeficiente también se puede expresar en función de las frecuencias absolutas:

$$K = \frac{F_c - F_a}{N - F_a} \qquad [5.6]$$

donde:

F_c = frecuencia observada de clasificaciones coincidentes.

F_a = frecuencia de coincidentes esperadas por azar.

N = número total de personas de la muestra.

Con los datos de la tabla 5.2,

	Test «B»		
Test «A»	Maestría	No- maestría	Total (N_i)
Maestría	2	3	5
No-maestría	1	14	15
Total (N_j)	3	17	N = 20

En primer lugar, se calcula el número de coincidencias esperadas por azar, esto se hace a partir de las frecuencias marginales.

$$\frac{3 \cdot 5}{20} = 0,75$$

$$\frac{17 \cdot 15}{20} = 12,75$$

$$F_a = 0,75 + 12,75 = 13,50$$

A continuación, se calcula el número de clasificaciones coincidentes observadas:

$$F_c = 2 + 14 = 16$$

Por lo tanto:

$$K = \frac{F_c - F_a}{N - F_a} = \frac{16 - 13,50}{20 - 13,50} = \frac{2,50}{6,50} = 0,38$$

Como se puede observar, el valor es el mismo que se ha obtenido anteriormente.

Para ver si el valor del coeficiente Kappa obtenido es estadísticamente significativo, Cohen (1960) propuso la utilización del error típico de medida de *K:*

$$S_e = \sqrt{\frac{F_a}{N(N - F_a)}} \qquad [5.7]$$

La hipótesis nula que se plantea es H_0: $K = 0$, y como hipótesis alternativa H_1: $K \neq 0$. En el caso de rechazar la *Ho* se puede establecer que el valor del coeficiente Kappa obtenido es estadísticamente significativo.

Aplicando los datos de nuestro ejemplo:

En primer lugar calculamos el error típico de medida

$$S_e = \sqrt{\frac{13,50}{20(20 - 13,50)}} = \sqrt{\frac{13,50}{130}} = 0,32$$

A continuación calculamos el intervalo confidencial:

$$K \pm Z_x \cdot S_e \qquad [5.8]$$

Si utilizamos un N.C. del 95%, el intervalo confidencial vendrá dado por:

$$0,38 \pm 1,96 \cdot 0,32 \Rightarrow -0,247 \leq K \leq 1$$

Dado que el valor $K = 0$, se encuentra dentro de los límites del intervalo, podemos establecer que el acuerdo entre las clasificaciones no es estadísticamente significativo.

5.1.3. Índice de Crocker y Algina

Crocker y Algina (1986) proponen el índice P^*, como una alternativa al coeficiente Kappa de Cohen (1960). Este índice se basa en que la probabilidad mínima de una decisión consistente es 0,50. Este mínimo tendrá lugar si las puntuaciones del test son estadísticamente independientes y el punto de corte está en la mediana de la distribución conjunta de las puntuaciones obtenidas por los sujetos en las dos aplicaciones. El coeficiente P^* viene expresado por:

$$P^* = \frac{p_c - 0,50}{1 - 0,50} = 2p_c - 1 \qquad [5.9]$$

Siguiendo a Crocker y Algina (1986), el valor de P^* es igual a 1 cuando las decisiones son totalmente consistentes, e igual a 0 cuando las decisiones no son más consistentes que las que resultarían de utilizar tests estadísticamente independientes, cuyas puntuaciones presentan la misma distribución y un punto de corte igual a la mediana de la distribución común.

En nuestro caso p_c =0,80, por lo tanto:

$$P^* = \frac{p_c - 0,50}{1 - 0,50} = 2p_c - 1 = (2 \cdot 0,80) - 1 = 0,60$$

5.2. Índices de acuerdo que requieren una sola aplicación del test

5.2.1. Método de Huynh

Los métodos que se han presentado implican la existencia de una sola muestra de sujetos y dos aplicaciones de un mismo test o de dos formas paralelas. Una de las primeras ventajas que supone el método de Huynh es que sólo se precisa un test y una sola aplicación. En el trabajo de Keats y Lord (1962): *A theoretical distribution for mental test scores* estos autores proponen un método para pronosticar las puntuaciones en un test «B» conocidas las puntuaciones de una muestra de sujetos en una primera aplicación (test «A»). El método original descrito por Huynh (1976) lleva consigo un desarrollo matemático laborioso por lo que es aconsejable que se cuente con paquetes de programas computerizados. El lector interesado puede seguir este desarrollo en Berk (1980).

Sin embargo, Huynh (1976) y Peng y Subkoviak (1980), han propuesto un método de aproximación más manejable. Esta aproximación al procedimiento anterior, presupone que la distribución de puntuaciones es aproximadamente normal. Huynh sugiere que este presupuesto es adecuado cuando el número de ítems es superior a ocho y la razón entre la media de las puntuaciones de los sujetos en el test y el número de ítems oscila entre 0,15 y 0,85. Los pasos a seguir según este método son:

1) Calcular la media ($\bar{X}$), la varianza (S_x^2) y el coeficiente de correlación de Kuder-Richardson 21 (*KR21*) y especificar el valor del punto de corte (C). En el ejemplo que presentamos, suponemos que la media del test A es igual a $\bar{X} = 5,15$, la varianza $S_x^2 = 4,45$, el coeficiente KR21= 0,37 y el punto de corte sobre las puntuaciones directas X se establece en $C = 7$.
2) Calcular la puntuación típica (Zx) correspondiente al valor del punto de corte, con una corrección de 0,5 y, acudiendo a las tablas de curva normal se busca el valor de P que deja por debajo la Z obtenida.

$$Z_x = \frac{(C - 0,5 - \bar{X})}{S_x} \qquad [5.10]$$

$$Z_x = \frac{(C - 0,5 - \bar{X})}{S_x} = \frac{(7 - 0,5 - 5,15)}{2,109} = 0,64$$

$$Z_x = 0,64 \rightarrow p_z = 0,74$$

3) A partir de las tablas de Gupta (1963) incluidas al final del libro (tabla 11), obtenemos la probabilidad (P_{zz}) de que dos variables distribuidas normalmente con una correlación KR21= 0,37 sean menores que $Z = 0,64$. (Se toma el valor por aproximación).

$$P_{zz} = 0,58$$

4) Por último calculamos los valores p_c y k

$$p_c = 1 + 2(p_{zz} - p_z) \qquad [5.11]$$

$$k = \frac{p_{zz} - p_z^2}{p_z - p_z^2} \qquad [5.12]$$

$$pc = 1 + 2(p_{zz} - p_z) = 1 + 2(0{,}58 - 0{,}74) = 0{,}68$$

$$k = \frac{p_{zz} - p_z^2}{p_z - p_z^2} = \frac{0{,}58 - (0{,}74)^2}{0{,}74 - (0{,}74)^2} = 0{,}16$$

En definitiva podemos considerar que el método de Huynh constituye un procedimiento matemático sofisticado para estimar la consistencia de clasificación a partir de una sola administración de un test de maestría (Subkoviak, 1980).

Nota: Téngase en cuenta que al utilizar la fórmula KR21, los ítems del test deberán tener la misma dificultad.

5.2.2. Método de Subkoviak

Subkoviak (1980) establece un procedimiento con una única aplicación cuando no es posible establecer una forma paralela de un test. El método de Subkoviak simula las puntuaciones de una segunda forma paralela del test. Su método, al igual que el desarrollado por Huynh, proporciona una buena estimación de los valores p_c y k.

Para la explicación del método de Subkoviak vamos a utilizar los datos del ejemplo desarrollado en el método de Hambleton y Novick suponiendo que solo se pudiera aplicar el test A y que el coeficiente de fiabilidad fuera igual a 0,62.

Los pasos para confeccionar la tabla son los siguientes:

— Las columnas 1, 2, 3 y 4 representan la distribución de frecuencias de las puntuaciones obtenidas por los 20 sujetos de la muestra.

— Una vez obtenida la distribución de frecuencias, se calcula la media y el coeficiente alfa del test, que suponemos igual a 0,62:

$$\bar{X} = \frac{\sum X}{N} = \frac{103}{20} = 5{,}15$$

$$\alpha = 0{,}62$$

— A continuación se estima la probabilidad de que una persona con una determinada puntuación X responda correctamente a cada ítem. Dicha probabilidad se estima mediante la ecuación:

$$p_x = \alpha\left(\frac{X}{n}\right) + (1-\alpha)\left(\frac{\bar{X}}{n}\right) \qquad [5.13]$$

donde:

α = coeficiente alfa.

X = Puntuación directa.

N = Número d ítems del test.

$\bar{X}$ = Media del test.

TABLA 5.3
Método de Subkoviak

X	f_x	p_x	P_x	$1\text{-}2\,(P_x - P_x^2)$	$f_x\,(1\text{-}2\,(P_x - P_x^2))$	$f_x P_x$
9	1	0,628	0,7362	0,6115	0,6115	0,7362
8	2	0,576	0,5999	0,5198	1,0396	1,1998
7	2	0,525	0,4562	0,5038	1,0076	0,9124
6	4	0,473	0,3164	0,5674	2,2696	1,2656
5	4	0,421	0,1978	0,6826	2,7304	0,7912
4	2	0,370	0,1105	0,8034	1,6068	0,2210
3	3	0,318	0,0522	0,9010	2,7030	0,1566
2	1	0,266	0,0201	0,9606	0,9606	0,0201
1	1	0,215	0,0059	0,9882	0,9882	0,0059
	20				13,9173	5,3089

A modo de ilustración, calcularemos el resultado para el primer caso de la matriz de frecuencias, es decir, el caso en el que $X = 9$. El resto de los valores de p_x (representados en la tercera columna) se obtienen siguiendo el mismo proceso. Recuérdese que el test consta de 12 ítems.

$$p_x = 0{,}62\ (9/12) + (1\text{–}0{,}62)\ (5{,}15/12) = 0{,}628$$

— En tercer lugar calculamos la probabilidad de que una persona, con una determinada puntuación X, y una probabilidad p_x de acertar cada ítem (valor correspondiente en la columna

3) responda correctamente siete o más ítems en el test y sea clasificado dentro del grupo de maestría. Para ello, puesto que podemos considerar los ítems como ensayos de un proceso binomial, aplicaremos la función de distribución binomial o se buscarán los valores correspondientes en las tablas de la distribución binomial, para lo que se tendrá en cuenta el número de ítems (n), el valor del punto de corte (c), que en nuestro ejemplo es 7 y la probabilidad de acertar cada ítem (p_x) en función de la puntuación obtenida en el test.

$$f(k) = \text{Prob}(X \geq k) = \sum \binom{n}{x} p^x q^{n-x} \qquad [5.14]$$

donde:

$$\sum \binom{n}{x} p^x q^{n-x} = \sum_x^n \left(\frac{n!}{x!(n-x)!} \right) p^x q^{n-x}$$

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots 2 \cdot 1$$

Los valores obtenidos aparecen recogidos en la cuarta columna de la tabla 5.3.

Veamos cuál sería el proceso a seguir en el caso de un sujeto que ha obtenido una puntuación 9 en el test, y una probabilidad de acertar cada ítem de $p_x = 0{,}628$. Recuérdese que el punto de corte se estableció en 7 ítems.

$$f(7) = \text{Prob}(X = 7) = \binom{12}{7} 0{,}628^7 0{,}372^5 = 0{,}21734$$

$$f(8) = \text{Prob}(X = 8) = \binom{12}{8} 0{,}628^8 0{,}372^4 = 0{,}22932$$

$$f(9) = \text{Prob}(X = 9) = \binom{12}{9} 0{,}628^9 0{,}372^3 = 0{,}17206$$

$$f(10) = \text{Prob}(X = 10) = \binom{12}{10} 0{,}628^{10} 0{,}372^2 = 0{,}087$$

$$f(11) = \text{Prob}(X = 11) = \binom{12}{11} 0{,}628^{11} 0{,}372^1 = 0{,}02675$$

$$f(12) = \text{Prob}(X = 12) = \binom{12}{12} 0{,}628^{12} 0{,}372^0 = 0{,}00376$$

$$P_x = 0{,}73623$$

Luego, la probabilidad de acertar 7 o más ítems de 12 es $P_x = 0{,}73623$ que es la suma de las probabilidades de acertar 7, 8, 9, 10, 11 y 12.

— Una vez calculados los valores de la cuarta columna, se calcula la probabilidad de que cada sujeto sea consistentemente clasificado en el grupo de maestría para dos tests independientes; es decir, la probabilidad de que cada persona sea clasificada en el grupo de maestría por el primer test (P_{x_1}), por la probabilidad de que sea clasificado en el grupo de maestría por el segundo test (P_{x_2}) y que será igual a P_x^2 y la probabilidad de que sea clasificado en el grupo de no maestría en los dos tests que será:

$$\left(1 - P_{x_1}\right)\left(1 - P_{x_2}\right) = \left(1 - P_x\right)^2 \qquad [5.15]$$

Consiguientemente, la probabilidad de clasificación consistente para dicho sujeto es:

$$P_x^2 + \left(1 - P_x\right)^2 = 1 - 2 \cdot \left(P_x - P_x^2\right) \qquad [5.16]$$

En nuestro caso tenemos que:

$$1 - 2 \cdot (0{,}7362 - 0{,}7362^2) = 0{,}6115$$

El conjunto de todos los valores obtenidos aparece recogido en la quinta columna.

— En la sexta columna se recoge el número de sujetos que, habiendo obtenido una puntuación X serán consistentemente clasificados. Para el caso de $X = 9$, tenemos:

$$f_x \cdot \left[1 - 2\left(P_x - P_x^2\right)\right] \qquad [5.17]$$

$$1 \cdot \left[1 - 2 \cdot (0{,}7362 - 0{,}7362^2)\right] = 0{,}6115$$

La forma de obtener estos valores es multiplicando los valores obtenidos en la quinta columna por la frecuencia de la columna 2.

— Por último, la suma de los valores de la columna 7, que se obtienen multiplicando los valores obtenidos en la columna 4 por los de la columna 2, representa el número de sujetos que superarán el punto de corte en ambos tests.

Con todos estos datos ya se pueden obtener los coeficientes p_c y *Kappa*

El coeficiente p_c se obtiene dividiendo el valor de la suma del número de sujetos que para una determinada puntuación han sido consistentemente clasificados (columna 6) por el número total de sujetos.

$$p_c = \frac{\sum f_x \left[1 - 2(P_x - P_x^2)\right]}{f_x} \qquad [5.18]$$

$$p_c = \frac{\sum f_x \left[1 - 2\left(P_x - P_x^2\right)\right]}{f_x} = \frac{13,9173}{20} = 0,696$$

En dicha expresión, el numerador representa el número de sujetos correctamente clasificados, y el denominador el número total de sujetos.

Para calcular el coeficiente *Kappa* hay que calcular el valor de la probabilidad de clasificación consistente por azar (p_a) a partir de la suma del número total estimado de sujetos clasificados en el grupo de maestría cuyos valores podemos ver en la columna 7.

$$p_a = 1 - 2\left(\frac{\sum f_x \cdot P_x}{N} - \left(\frac{\sum f_x \cdot P_x}{N}\right)^2\right) \qquad [5.19]$$

$$p_a = 1 - 2\left(\frac{\sum f_x \cdot P_x}{N} - \left(\frac{\sum f_x \cdot P_x}{N}\right)^2\right) = 1 - 2\left(\frac{5,3089}{20} - \left(\frac{5,3089}{20}\right)^2\right) = 0,61$$

A continuación calculamos el coeficiente Kappa:

$$K = \frac{p_c - p_a}{1 - p_a} \qquad [5.20]$$

$$K = \frac{p_c - p_a}{1 - p_a} = \frac{0,696 - 0,61}{1 - 0,61} = \frac{0,086}{0,39} = 0,22$$

5.2.3. Coeficiente de Livingston

El coeficiente de Livingston (1972) se desarrolla en el contexto de la Teoría Clásica de los Tests. Siguiendo a Muñiz (1998), podemos decir que los métodos que hemos presentado hasta el momento para el estudio de la fiabilidad, consideran, por igual, tanto los errores que cometemos cuando clasificamos a un sujeto perteneciente al grupo de maestría en el grupo de no-maestría, como los que cometemos cuando clasificamos a un sujeto perteneciente al grupo de no-maestría dentro del grupo de maestría. Sin embargo, el coeficiente de Livingston sí tiene en cuenta este tipo de errores, al considerar más importantes los errores de clasificación de los sujetos más distanciados del punto de corte de aquellos que están más cerca del punto de corte. Lógicamente, es más fácil cometer errores de clasificación cuando un sujeto se encuentra muy cercano al punto de corte y será más difícil cometer estos errores de clasificación cuando el sujeto se encuentra muy alejado del punto de corte.

El coeficiente viene determinado por:

$$K_{xv}^2 = \frac{\alpha \cdot S_x^2 + (\bar{X} - C)^2}{S_x^2 + (\bar{X} - C)^2} \qquad [5.21]$$

donde:

α = coeficiente alfa.

S_x^2 = varianza del test.

$\bar{X}$ = media del test.

C = punto de corte.

EJEMPLO:

Si aplicamos la fórmula a los datos del ejemplo anterior: $\alpha = 0,62$, $\bar{X} = 5,15$ y $S_x = 2,109$ y el punto de corte igual a 7:

$$K_{xv}^2 = \frac{\alpha \cdot S_x^2 + (\bar{X} - C)^2}{S_x^2 + (\bar{X} - C)^2} = \frac{0,62 \cdot 4,45 + (5,15 - 7)^2}{4,45 + (5,15 - 7)^2} = \frac{2,759 + 3,42}{7,87} = 0,78$$

A medida que el punto de corte se distancia del valor de la media del test, aumenta el valor de K_{xv}^2. Cuando la media del test coincide con el punto de corte, K_{xv}^2 es igual al coeficiente alfa. Cuando el coeficiente de fiabilidad alfa es igual a 1, K_{xv}^2 también es igual a uno. Por lo tanto, K_{xv}^2 será siempre igual o mayor que el coeficiente de fiabilidad alfa.

6. MÉTODOS PARA ESTIMAR EL PUNTO DE CORTE EN LOS TESTS REFERIDOS AL CRITERIO

En el punto anterior hemos presentado una serie de métodos para el cálculo de la fiabilidad de los tests referidos al criterio en los cuales partimos del establecimiento de una puntuación de corte que nos va a permitir clasificar a un sujeto en dos posibles categorías: la de aquellos sujetos que dominan el criterio evaluado, o la de aquellos sujetos que no dominan el criterio evaluado. Es decir, el criterio actúa como un filtro o punto de corte para clasificar a los sujetos. La cuestión fundamental es, ¿cómo se establece este punto de corte? ¿cuál es la puntuación a partir de la cual un sujeto se situará en un grupo u otro? Existen innumerables situaciones que requieren establecer un punto de corte antes de dotar de significado a la puntuación obtenida por un sujeto en un test. Por ejemplo, la calificación de aprobado o suspenso en un examen, la selección de aspirantes a un puesto de trabajo, la admisión para entrar en la universidad, son ejemplos donde es necesario establecer un punto de corte. Como se puede observar, las decisiones que se tomen como consecuencia del valor del punto de corte establecido son de gran importancia, ya que de ellas dependerá, en algunos casos, el futuro de las personas implicadas.

Por lo general, se suele contar con un número adecuado de expertos que son quienes establecen ese punto de corte. Es, en definitiva, una cuestión sujeta a un grado de subjetividad, por lo que una garantía absoluta no existe cuando se establece dicho punto de corte. Siempre habrá sujetos clasificados erróneamente. Sujetos clasificados como competentes cuando no lo son y viceversa.

Se suelen considerar dos tipos de puntos de corte (Muñiz, 1998): *puntos de corte relativos y puntos de corte absolutos*. Se definen como relativos, cuando el punto de corte se establece en función del grupo de sujetos evaluados, y se definen como absolutos, cuando el punto de corte se establece en función del constructo o materia objeto de estudio.

Son innumerables los modelos propuestos (Berk, 1996, 1986; Cizek, 1996; Hambleton y Eignor, 1980; Hambleton y Rogers, 1990; Jaeger, 1995, 1989) para establecer el punto de corte. Aquí presentamos los métodos utilizados con mayor frecuencia.

6.1. Métodos valorativos

Los cuatro métodos que veremos a continuación se basan en la evaluación que un grupo de expertos, con un cierto entrenamiento y en número suficiente, llevan a cabo sobre los ítems de un test. La forma en que dichos expertos abordan la evaluación también varía según el método utilizado. Los expertos solamente deben ser especialistas en la materia a evaluar, y no es necesario que conozcan el grado de competencia de cada uno de los sujetos. A pesar de que aquí solamente

presentaremos los modelos basados en el contenido de los ítems, existen otros modelos que basan el proceso de evaluación en el contenido del test (Glass, 1978; Shepard, 1976), o en características tales como el acierto al azar (Millman, 1973).

Método de Nedelsky

El método de Nedelsky (1954) es el primero de los procedimientos establecidos para fijar el punto de corte en tests de competencia mínima. Estos tests se utilizan habitualmente en el ámbito académico para determinar si un sujeto posee los conocimientos mínimos exigibles en una determinada materia. El método de Nedelsky se utiliza con tests compuestos de ítems de elección múltiple, y precisa que los expertos o jueces analicen las distintas alternativas de los ítems y, a continuación, determinen cuáles de las posibles alternativas serán consideradas como erróneas por un sujeto que tuviese los conocimientos mínimos exigibles para ser considerado como competente. El modelo asume que un sujeto elegirá al azar, entre las restantes opciones, la posible respuesta correcta.

Seguidamente, para cada ítem, el juez registra el recíproco del número de preguntas que quedan. Supongamos que un ítem consta de seis alternativas, y un juez considera que un sujeto mínimamente competente rechazará cuatro de ellas como erróneas. El recíproco, se determina dividiendo la unidad por el número de alternativas restantes, las que el sujeto no ha considerado como alternativas erróneas, en nuestro caso 2 por lo que el recíproco será 0,5. Esta puntuación se correspondería con la puntuación esperada para un sujeto en un ítem determinado. Para calcular la puntuación de un sujeto mínimamente cualificado en un test, se sumarían todos los valores esperados de cada ítem. De esta manera, se obtendrá la puntuación otorgada por un determinado juez a un sujeto mínimamente cualificado. El promedio de las puntuaciones otorgadas por todos los jueces, nos dará la puntuación de corte.

Veamos el proceso que se seguiría con el siguiente ítem correspondiente a un test de mecánica:

Una pieza esencial para que un vehículo pueda circular es:

a) El manillar

b) El espejo retrovisor

c) El motor de arranque

d) La rueda de repuesto

e) Los intermitentes

f) Los faros

Según el método de Nedelsky, un juez consideraría que un sujeto, con unos conocimientos mínimos de mecánica descartaría como alternativas erróneas la *a, b y d*. La puntuación esperada para un sujeto mínimamente competente en ese ítem vendría dada por el resultado de dividir la unidad entre el número de alternativas que se supone que el sujeto no ha rechazado como erróneas; en nuestro caso 1: 3 = 0,33. Este proceso es el que se seguiría con todos los ítems del test. El valor esperado por ese juez para ese tipo de sujeto en el test será igual a la suma de los valores esperados en cada uno de los ítems. Si se calcula la media de todos los valores esperados por todos los jueces se tendrá el valor del punto de corte.

Para corregir los posibles efectos del azar a la hora de determinar el punto de corte se puede utilizar la siguiente expresión:

$$P_c = A - \frac{N - A}{n - 1} \qquad [5.22]$$

donde:

P_c = la puntuación corregida.

N = número de ítems.

A = media de los valores esperados.

n = número de alternativas de cada ítem.

EJEMPLO:

Supongamos un test de percepción del color compuesto por 40 ítems de 4 alternativas. La media de los valores esperados determinada por 7 jueces es 28. Esto implica que el valor del punto de corte sin corregir el azar es igual a 28. Veamos cuál sería el valor si corregimos los efectos del azar.

$$P_c = A - \frac{N - A}{n - 1} = 28 - \frac{40 - 28}{4 - 1} = 28 - \frac{12}{3} = 24$$

Corregido el efecto del azar la puntuación de corte sería 24.

A pesar de su utilización no deja de ser un procedimiento cuestionable. El método de Nedelsky asume que los sujetos responden al azar entre las alternativas que no son descartadas como erróneas cuando no conocen la respuesta correcta, sin embargo, no existe ninguna evidencia que sustente este hecho (van der Linden, 1982; Jaeger, 1989). Asimismo es un método en el que se tiende a dar valores de corte más bajos que si se utilizan otros procedimientos (Shepard, 1980) debido a que los jueces no suelen asignar valores esperados entre 0,5 y 1. De ser así, o sólo quedarían dos alternativas sin eliminar y, por lo tanto, el valor esperado sería 0,5 o sólo quedaría una alternativa sin eliminar, en cuyo caso el valor esperado sería 1.

Método de Angoff

El método propuesto por Angoff (1971), puede considerarse como una variante del método de Nedelsky, con la diferencia de que es aplicable a toda clase de ítems, no sólo a los de elección múltiple. En este método, no se pide a los jueces que emitan juicios acerca de cada una de las alternativas de un ítem, como en el método anterior, sino que deben evaluar el ítem globalmente y determinar la probabilidad de que un sujeto, con los requisitos mínimos para ser competente, responda correctamente a cada uno de los ítems del test. Para poder determinar estas probabilidades, los jueces han de comprender claramente la tarea que deben realizar los sujetos. Una vez que los distintos jueces han establecido las probabilidades de que los sujetos mínimamente competentes respondan a los ítems correctamente, estamos en condiciones de establecer el punto de corte. La puntuación total establecida por cada uno de los jueces para cada sujeto se considera como la puntuación estimada de un sujeto mínimamente competente. Para calcular el punto de corte, se suman los valores de las probabilidades establecidas por cada uno de los jueces, y se calcula la media de dichas puntuaciones. Como en el caso del método de Nedelsky, también se puede aplicar la corrección de los efectos del azar.

EJEMPLO:

En la tabla siguiente aparecen las probabilidades, otorgadas por cuatro jueces, de que un sujeto mínimamente competente supere cada uno de los ítems de un test. Calcular el punto de corte mediante el método de Angoff.

Ítems	Juez 1	Juez 2	Juez 3	Juez 4
1	0,44	0,25	0,45	0,20
2	0,35	0,20	0,40	0,38
3	0,40	0,25	0,35	0,30
4	0,30	0,40	0,30	0,45
5	0,50	0,22	0,50	0,25
6	0,30	0,30	0,45	0,22
Total	2,29	1,62	2,45	1,80

El punto de corte será igual a la media de las puntuaciones totales otorgadas por los cuatro jueces:

$$P.C. = \frac{2,29 + 1,62 + 2,45 + 1,80}{4} = 2,04$$

Método de Ebel

El método de Ebel (1972) guarda una cierta similitud con el método de Angoff que acabamos de ver, puesto que los jueces también realizan una valoración global del ítem aunque desde una doble perspectiva. Los jueces evalúan el grado de dificultad del ítem, y también su grado de relevancia. Ebel sugiere tres niveles de dificultad para cada ítem: *fácil, medio y difícil*; y cuatro niveles de relevancia: *esencial, importante, aceptable y dudoso*. De esta manera, se obtiene una matriz con doce categorías distintas en la que aparecerán clasificados todos los ítems del test. Una vez que se han clasificado los ítems en la casilla correspondiente, se hace un recuento del número de ítems por casilla, y los distintos jueces proceden a establecer un porcentaje que representa el número de ítems que serían contestados correctamente por un sujeto con una competencia mínima. A continuación se calcula el punto de corte mediante la siguiente ecuación:

$$X_c = \Sigma\, p(M) \qquad [5.23]$$

donde:

X_c = puntuación correspondiente al punto de corte.

p = proporción de ítems en cada casilla que un sujeto mínimamente competente debería contestar correctamente.

M = número de ítems en cada celda.

EJEMPLO:

En la siguiente tabla aparecen clasificados los 175 ítems de un test y el porcentaje de ítems de cada casilla que un juez considera que responderá correctamente un sujeto mínimamente competente (dividido por 100 se obtendrá la proporción). Calcular el punto de corte.

	Niveles de dificultad		
Niveles de relevancia	**Fácil**	**Medio**	**Difícil**
Esencial	Ítems: 15 Juez: 80%	Ítems: 20 Juez: 60%	Ítems: 10 Juez: 30%
Importante	Ítems: 30 Juez: 70%	Ítems: 18 Juez: 55%	Ítems: 7 Juez: 30%
Aceptable	Ítems: 25 Juez: 65%	Ítems: 15 Juez: 50%	Ítems:10 Juez: 25%
Dudoso	Ítems: 14 Juez: 40%	Ítems: 6 Juez: 45%	Ítems:5 Juez: 20%

$$X_c = \Sigma p(M) = 15\ (0,80) + 20\ (0,60) + 10\ (0,30) + 30\ (0,70) + 18\ (0,55) + \\ + 7\ (0,30) + 25\ (0,65) + 15\ (0,50) + 10\ (0,25) + 14\ (0,40) + 6\ (0,45) + \\ + 5\ (0,20) = 95,55$$

Esta puntuación correspondería, tal y como hemos expuesto a la puntuación otorgada por un juez; en el caso de que hubiera varios jueces, el valor del punto de corte vendría dado por la media de las puntuaciones asignadas por cada uno de ellos.

Método de Jaeger

El método propuesto por Jaeger (1978), puede considerarse una variante del método de Angoff. En este método se le pregunta a cada uno de los jueces, si cada uno de los ítems del test será contestado correctamente por los sujetos. El proceso para poder determinar el punto de corte precisa de tres sesiones. En la primera sesión, cada uno de los jueces, y para cada uno de los ítems del test, responde con un *Sí* o con un *No* a la pregunta de si un sujeto mínimamente competente será capaz de contestar correctamente ese ítem. Una vez que los jueces han contestado a dicha pregunta para cada uno de los ítems, se calcula el número de ítems a los que cada juez respondió con un *Sí*.

En la siguiente matriz se presentan los datos correspondientes a la evaluación que cinco jueces han hecho respecto a los siete ítems de un test.

En la segunda sesión, repetimos el mismo proceso que acabamos de describir pero, al comienzo de la sesión se pone a disposición de los jueces los datos obtenidos en la sesión anterior, las opiniones o recomendaciones emitidas por los jueces, y una tabla con los porcentajes de respuestas SI a cada uno de los ítems.

	Juez 1	Juez 2	Juez 3	Juez 4	Juez 5
Ítem 1	SI	SI	NO	SI	SI
Ítem 2	SI	NO	SI	SI	SI
Ítem 3	NO	NO	NO	SI	SI
Ítem 4	SI	NO	NO	SI	SI
Ítem 5	NO	NO	SI	NO	SI
Ítem 6	NO	NO	NO	NO	NO
Ítem 7	NO	NO	NO	NO	NO
Total	3	1	2	4	5

Una vez conocidos los datos de la primera sesión cada juez puede cambiar su opinión; en nuestro caso supongamos que obtenemos los siguientes resultados:

	Juez 1	Juez 2	Juez 3	Juez 4	Juez 5
Total	4	2	1	5	6

En la tercera sesión, se presentan los datos de la sesión anterior a los jueces, y se les pide que valoren nuevamente cada uno de los ítems. Al igual que en la sesión anterior los jueces pueden ir modificando sus juicios en función de la información que se les va proporcionando.

Una vez conocidos los datos de la segunda sesión supongamos que obtenemos los siguientes resultados:

	Juez 1	Juez 2	Juez 3	Juez 4	Juez 5
Total	4	3	5	6	7

El punto de corte, es la mediana más baja de los diferentes grupos de jueces.

Con los datos anteriores obtenemos los siguientes valores:

$Md_1 = 3$, $Md_2 = 4$, $Md_3 = 5$. Con estos resultados establecemos que el punto de corte es igual a 3.

Un problema de este método (Berk, 1986) es que sólo se permite la asignación de probabilidades de 0 ó 1, pues un sujeto o acierta o falla el ítem.

6.2. Métodos combinados

Los dos métodos que presentamos a continuación se basan en los juicios que los expertos llevan a cabo respecto a la competencia de los sujetos. En los métodos descritos en el apartado anterior, los jueces se suponían expertos en cuanto a los contenidos a evaluar. En los que ahora presentamos, además de esa condición, los jueces también deben conocer la competencia de los sujetos en la materia que se evalúa.

Método del grupo límite

En el método del grupo límite, propuesto por Zieky y Livingston (1977), se pide a los jueces que definan de mutuo acuerdo tres niveles de competencia en el dominio a evaluar: *competente, límite y no competente*. Seguidamente, los jueces deben identificar entre los sujetos a los que va dirigido el test, aquellos que, en su opinión, estarían en el límite de ser competentes. Es decir, aquellos sujetos cuyos conocimientos en la variable estudiada no son del todo inadecuados, pero tampoco adecuados como para ser considerados como competentes. Una vez que se han seleccionado los sujetos con estas características, se les aplica el test para, posteriormente, determinar el punto de corte. Para establecer el punto de corte, se calculará la media o la mediana de las puntuaciones que han obtenido en el test los sujetos *límite*. La mediana es más conveniente, puesto que es menos sensible a la variabilidad de las puntuaciones.

EJEMPLO:

Supongamos que una empresa conservera ha impartido a un grupo de trabajadores un cursillo de técnicas de envasado y etiquetado con el fin de poder aumentar sus ventas. Una vez terminado el cursillo, la dirección solicita de los técnicos que lo han impartido que emitan un juicio sobre el grado de aprovechamiento de quienes lo han realizado, observando que 7 de ellos parecen haber adquirido una formación límite. Una vez que los asistentes han sido sometidos a una prueba sobre adquisición de conocimientos, las puntuaciones de estos 7 sujetos fueron: 50, 48, 47, 46, 45, 43, 40.

Para calcular el punto de corte se podría calcular la media, aunque tal y como hemos apuntado es mejor calcular la mediana de estas puntuaciones que es igual a 46. Ese sería el punto de corte.

Método de los grupos de contraste

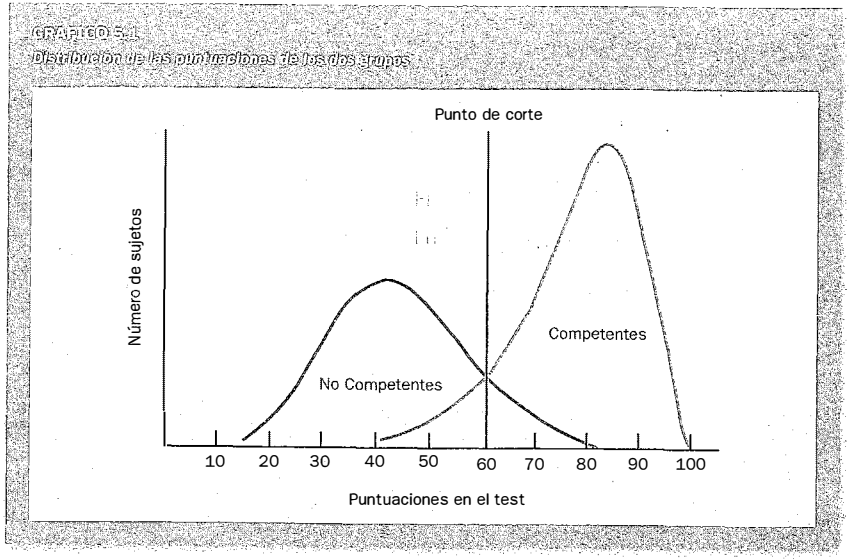
El método de los grupos de contraste (Berk, 1976; Livingston y Zieky, 1982), se basa, al igual que el método anterior, en el conocimiento que los jueces tienen del rendimiento de los sujetos en el dominio que se pretende evaluar con el test en el que estamos interesados en establecer el punto de corte. Una vez que los jueces han clasificado a los sujetos en dos grupos, los que a su juicio son competentes y los que no lo son, se les administra el test y las puntuaciones se establecen en base a su rendimiento en el mismo. El paso siguiente sería determinar el punto de corte. Para

ello se puede utilizar un procedimiento muy sencillo basado en la representación gráfica de la distribución de puntuaciones del grupo de sujetos considerados como competentes por los jueces, y la distribución de los que no son considerados como competentes (gráfico 5.1)

Se elegiría como punto de corte la intersección de ambas distribuciones, que en nuestro caso sería 60.

Si se desplaza el punto de corte hacia la derecha, se reducen los falsos positivos, es decir, se reduce la probabilidad de considerar como competentes a sujetos que no lo son. Por otra parte, si el punto de corte se desplaza hacia la izquierda, se reducen los falsos negativos, es decir, se reduce la probabilidad de considerar no competentes a los sujetos que sí lo son. Es fundamental tener en cuenta esto, ya que pueden surgir situaciones prácticas en las cuales puede interesar minimizar un tipo de error más que otro (Muñiz, 1998).

GRÁFICO 5.1
Distribución de las puntuaciones de los dos grupos



6.3. Métodos de compromiso

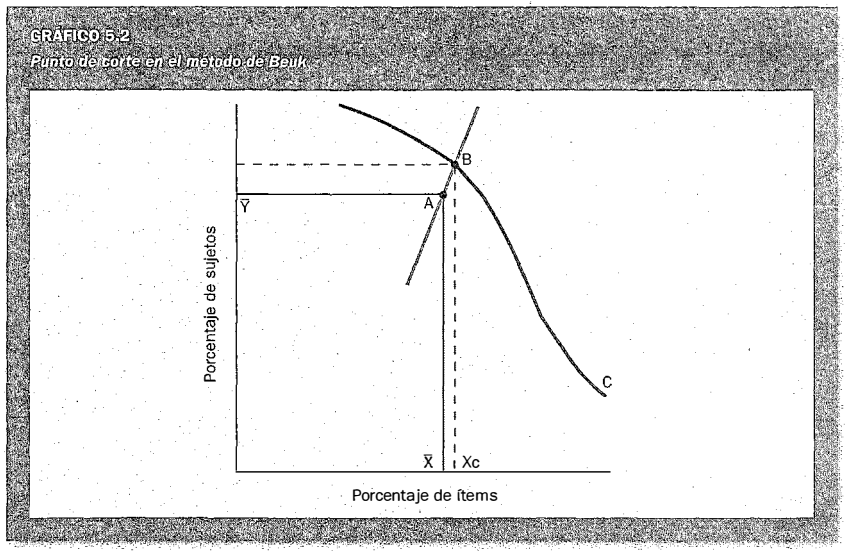
En los dos métodos que exponemos a continuación, el método de Beuk y el método de Hosftee, los jueces no se basan exclusivamente, como hasta ahora, en los conocimientos mínimos que un sujeto tiene que poseer para superar el criterio, sino que incorporan además la información relativa a la posición de un sujeto con relación a su grupo. El hecho de considerar la información derivada de la posición que un sujeto puede ocupar respecto a su grupo, viene justificada por las implicaciones de carácter social, económico, etc. que, en ocasiones, se pueden derivar del establecimiento del punto de corte.

Método de Beuk

En el método propuesto por Beuk (1984), los jueces han de tener en cuenta, en primer lugar, las puntuaciones obtenidas por los sujetos en el test en el que estamos interesados en establecer un punto de corte y, en segundo lugar, la información recogida de las respuestas de los jueces a dos preguntas que les son formuladas. La primera pregunta hace referencia al porcentaje mínimo de ítems, que los distintos jueces creen que un sujeto debería contestar correctamente para superar el test y la segunda, hace referencia al porcentaje de sujetos que estiman que obtendrán la puntuación mínima para superar el test. La primera pregunta hace referencia a datos absolutos, es decir, a la información derivada del simple conocimiento de un sujeto con relación al valor del punto de corte. La segunda pregunta, implica información o cuestiones de carácter relativo, es decir, cuestiones que pueden tener una importancia económica, social, etc. y que no dependen exclusivamente del conocimiento que tenga un sujeto.

Una vez que los jueces han recogido y analizado esta información, se procede a determinar el punto de corte. A continuación, se expone la forma de obtener el punto de corte según el modelo de Beuk. La siguiente representación gráfica ha sido tomada de Beuk (1984).

GRÁFICO 5.2
Punto de corte en el método de Beuk



En primer lugar, se representa sobre el eje de abscisas el porcentaje mínimo de ítems que los distintos jueces creen que un sujeto debería contestar correctamente para superar el test y, en el eje de ordenadas, el porcentaje de sujetos que estiman que obtendrán la puntuación mínima para superar el test. A continuación calculamos el valor de la media de los juicios emitidos por los expertos a las dos preguntas formuladas, ($\overline{X}$ y $\overline{Y}$) y se representa el punto de intersección «A».

En segundo lugar, se obtiene la distribución «C» correspondiente a las puntuaciones de los sujetos en el test. Como se puede observar la distribución es decreciente ya que, a medida que el número de ítems que hay que responder correctamente para superar el test se eleva, disminuye el número de sujetos que lo superan.

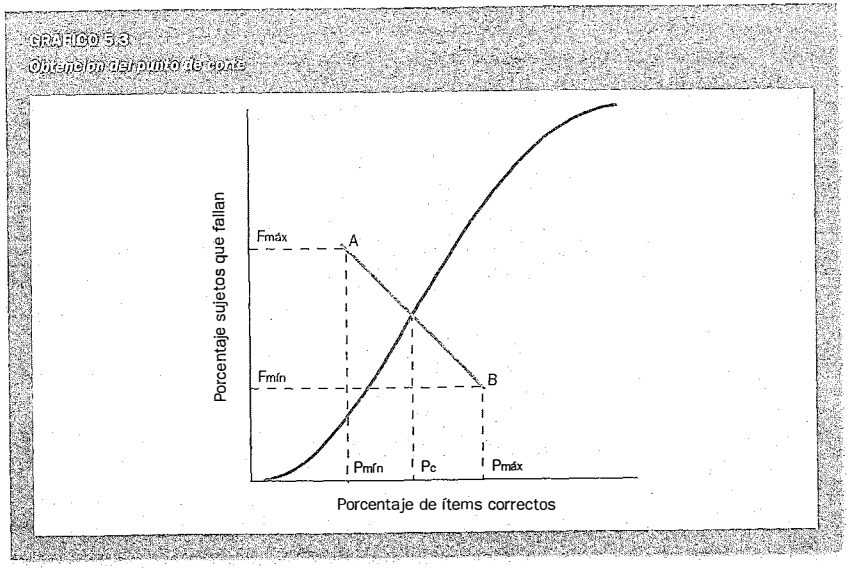
En tercer lugar, se dibuja una recta (AB), cuya pendiente viene determinada por el cociente entre la desviación típica de las respuestas emitidas por los jueces a las dos primeras preguntas: Sy/Sx.

En cuarto lugar, se obtiene el punto de corte X_c. Para obtener el punto de corte, se proyecta el punto «B» sobre el eje de abscisas. El punto de intersección determina el punto de corte X_c. Este valor expresa el porcentaje de ítems que un sujeto debe contestar correctamente. Si queremos expresar este valor en función del número de ítems, multiplicamos el valor de X_c por el número de ítems del test «n», es decir: Nº ítems = $X_c \cdot n$.

Método de Hofstee

El método de Hofstee (Hofstee, 1983; De Gruijter, 1985), se basa en la información proporcionada por los jueces al dar respuesta a cuatro puntos: el punto de corte que los jueces consideran adecuado y que se define como el porcentaje de ítems que los sujetos deben superar, el punto de corte que los jueces consideran inadecuado, el porcentaje máximo admisible de sujetos que fallan en el test y, el porcentaje mínimo admisible de sujetos que fallan en el test. Con esta información y la distribución de los resultados obtenidos en el test, se puede establecer el punto de corte mediante la siguiente representación gráfica 5.3 (Tomado de Muñiz, 1998):

GRÁFICO 5.3
Obtención del punto de corte



Para la obtención del punto de corte se procede de la siguiente manera: En primer lugar se representa en el eje de abscisas los puntos $P_{máx}$, punto de corte que los jueces consideran adecuado y $P_{mín}$, punto de corte que los jueces consideran inadecuado. En el eje de ordenadas se representan los puntos $F_{máx}$, porcentaje máximo admisible de sujetos que fallan en el test y $F_{mín}$, porcentaje mínimo admisible de sujetos que fallan en el test. A continuación se representan los puntos *A y B,* resultantes de las intersecciones $P_{mín}$ - $F_{máx}$ y $P_{máx}$ - $F_{mín}$, respectivamente. Por último se traza una recta perpendicular al eje de abscisas que coincida con la intersección de la distribución de las puntuaciones en el test, y la recta *AB* y se determina el punto P_c, punto de corte que buscamos.

7. EJERCICIOS DE AUTOEVALUACIÓN

1. Se han aplicado dos tests compuestos de 15 ítems a una muestra de 12 sujetos. Para que un sujeto sea clasificado dentro del grupo de maestría debe contestar correctamente un mínimo de 10 ítems. Calcular el índice de fiabilidad empleando para ello el método propuesto por Hambleton y Novick y el índice Kappa de Cohen.

SUJETOS	TEST - A	TEST - B
1	10	9
2	8	9
3	11	10
4	12	10
5	7	7
6	10	10
7	9	8
8	11	10
9	10	10
10	8	6
11	10	11
12	11	7

2. Calcular la probabilidad de que un sujeto sea clasificado dentro de un grupo de maestría, supuesta una puntuación de corte del 80%, $n = 10$, $x = 8$, $p = 0{,}75$.

3. En la matriz da datos adjunta se presenta la puntuación total obtenida por 10 sujetos en dos tests paralelos de fluidez verbal compuestos por diez ítems. Para que un sujeto sea clasificado dentro del grupo de maestría debe responder correctamente a un mínimo de 6 ítems.

	Puntuación total			Puntuación total	
Sujetos	Test «A»	Test «B»	Sujetos	Test «A»	Test «B»
1	7	6	6	7	8
2	9	8	7	5	5
3	8	9	8	8	7
4	5	6	9	6	5
5	3	4	10	7	9

Estimar la fiabilidad en las clasificaciones utilizando el coeficiente kappa de Cohen.

4. En la tabla adjunta se presentan las puntuaciones y frecuencias obtenidas por 25 sujetos en un test compuesto por 10 ítems. Para que un sujeto sea clasificado dentro del grupo de maestría, debe responder un mínimo de 8 ítems. Calcular, empleando el método de Subkoviak, la consistencia de clasificación una vez eliminada la proporción de clasificación debida al azar. (KR20 = 0,56)

X	f_x
9	1
8	2
7	3
6	3
5	5
4	6
3	3
2	1
1	1

5. En la siguiente tabla se presentan las probabilidades asignadas por tres jueces de que los cinco ítems de un test utilizado en un proceso de selección sean superados por un grupo de sujetos.

Ítem	Juez 1	Juez 2	Juez 3
1	0,7	0,8	0,9
2	0,8	0,7	0,8
3	0,5	0,6	0,7
4	0,4	0,5	0,5
5	0,4	0,3	0,4

Calcular:

a. Los puntos de corte de cada Juez mediante el método de Angoff.

b. El punto de corte del test, a partir de la información de los tres Jueces.

c. Qué Juez considera el test más fácil y más difícil.

6. Hemos aplicado un test de aptitud numérica a un grupo de estudiantes de 1º de Bachillerato. El test está compuesto por ítems de elección múltiple con cuatro posibles alternativas. En la siguiente tabla se recogen las alternativas erróneas que cuatro jueces creen que serían descartadas por un alumno con los conocimientos mínimos exigidos para superar el test.

Ítem	Juez 1	Juez 2	Juez 3	Juez 4
1	*bcd*	*cd*	*bc*	*bcd*
2	*cd*	*bd*	*bcd*	*cb*
3	*ab*	*abd*	*abd*	*bd*
4	*acd*	*ac*	*cd*	*acd*

Calcular:

a. El valor esperado en el test para cada Juez.

b. El punto de corte del test sin corregir y corrigiendo el efecto azar, utilizando el método de Nedelsky

7. Ejercicios conceptuales

Ante cada una de las afirmaciones que se muestran a continuación, el lector deberá responder si el concepto que contiene es verdadero o falso.

1. El coeficiente kappa (*K*) es un estimador de la consistencia de clasificación de sujetos.
2. El método propuesto por Subkoviak para determinar la fiabilidad en las clasificaciones requiere dos aplicaciones del test.
3. Los tests referidos al criterio evalúan la posición de un sujeto en función de su nivel de rendimiento respecto al dominio definido.
4. El valor del coeficiente Kappa oscila entre 0 y 1.
5. Si $p \geq p_c$, podemos establecer que un sujeto pertenece al grupo de maestría.
6. Un error falso-negativo tiene lugar cuando clasificamos incorrectamente a un sujeto dentro de un grupo de maestría.
7. La clasificación de un sujeto dentro de un grupo de maestría depende del valor p_c establecido.
8. El valor de kappa proporciona una medida de la consistencia de clasificación de los sujetos dependiente del valor esperado por azar.
9. En los tests referidos a la norma no se hace hincapié en la especificación clara del dominio de contenidos.

10. El índice *P** de Crocker y Algina se basa en el modelo binomial.
11. Los puntos de corte absolutos se establecen en función del grupo de sujetos evaluados.
12. El método de Angoff puede ser considerado como una variante del método de Nedelsky.
13. El método de Beuk es un método valorativo.

8. SOLUCIONES A LOS EJERCICIOS DE AUTOEVALUACIÓN

1.

	Test «B»		
Test «A»	Maestría	No- maestría	Total
Maestría	6	2	8
No-maestría	0	4	4
Total	6	6	12

$$p_c = \frac{6}{12} + \frac{4}{12} = 0,83$$

$$p_a = \frac{8}{12} \cdot \frac{6}{12} + \frac{4}{12} \cdot \frac{6}{12} = 0,50$$

$$k = \frac{p_c - p_a}{1 - p_a} = \frac{0,83 - 0,50}{1 - 0,50} = \frac{0,33}{0,50} = 0,66$$

2. Puntuación de corte del 80%, $n = 10$, $x = 8$ y, $p = 0.75$

Aplicando la función de distribución binomial:

$$\text{Prob}(x \geq 8 | p = 0,75, n = 10) = \sum_{x=8}^{10} \binom{10}{x} \cdot (0,75)^x \cdot (0,25)^{n-x}$$

$$\text{Prob}(x = 8) = \binom{10}{8} \cdot (0,75)^8 \cdot (0,25)^2 = 45 \cdot 0,10 \cdot 0,0625 = \quad 0,28$$

$$\text{Prob}(x = 9) = \binom{10}{9} \cdot (0,75)^9 \cdot (0,25)^1 = 10 \cdot 0,075 \cdot 0,25 = \quad 0,19$$

$$\text{Prob}(x = 10) = \binom{10}{10} \cdot (0,75)^{10} \cdot (0,25)^0 = 1 \cdot 0,056 \cdot 1 = \quad 0,056$$

$$\Sigma = 0,53$$

La probabilidad de acertar 8 o más ítems de 10 y ser clasificado dentro del grupo de maestría es igual a 0,53.

3. $N = 10 \quad n = 10$

Puntuación Total			Puntuación Total		
Sujetos	**Test «A»**	**Test «B»**	**Sujetos**	**Test «A»**	**Test «B»**
1	7	6	6	7	8
2	9	8	7	5	5
3	8	9	8	8	7
4	5	6	9	6	5
5	3	4	10	7	9

	Test «B»		
Test «A»	Maestría	No- maestría	Total
Maestría	6	1	7
No-maestría	1	2	3
Total	7	3	10

Se calculan las frecuencias de coincidencias esperadas por azar:

$$\frac{7 \cdot 7}{10} = 4,90$$

$$F_a = 4,90 + 0,90 = 5,80$$

$$\frac{3 \cdot 3}{10} = 0,90$$

A continuación, calculamos las frecuencias observadas de clasificaciones coincidentes

$$F_c = 6 + 2 = 8$$

Por lo tanto:

$$k = \frac{F_c - F_a}{N - F_a} = \frac{8 - 5,80}{10 - 5,80} = \frac{2,20}{4,20} = 0,52$$

Este resultado nos indica una consistencia de clasificaciones media.

4.

X	f_x	p_x	P_x	$1\text{-}2\,(P_x - P_x^2)$	$f_x\,(1\text{-}2\,(P_x - P_x^2))$	$f_x P_x$
9	1	0,724	0,4486	0,5052	0,5052	0,4492
8	2	0,668	0,3023	0,5782	1,1564	0,6045
7	3	0,612	0,1874	0,6954	2,0862	0,5623
6	3	0,556	0,1064	0,8099	2,4296	0,3192
5	5	0,500	0,0547	0,8966	4,4830	0,2734
4	6	0,444	0,0250	0,9512	5,7071	0,1502
3	3	0,388	0,0100	0,9803	2,9409	0,0299
2	1	0,332	0,0033	0,9934	0,9934	0,0033
1	1	0,276	0,0009	0,9983	0,9983	0,0009
	25				21,3001	2,3929

$$\overline{X} = \frac{125}{25} = 5$$

Veamos como se han obtenido los valores de p_x y P_x para el caso de $X = 9$

$$p_x = 0{,}56\ (9/10) + (1\text{–}0{,}56)\ (5/10) = 0{,}724$$

Aplicando la función de distribución binomial:

$$\text{Prob}(x \geq 8 | p = 0{,}56, n = 10) = \sum \binom{10}{x} \cdot (0{,}724)^x \cdot (0{,}276)^{n-x}$$

$$\text{Prob}(x = 8) = \binom{10}{8} \cdot (0{,}724)^8 \cdot (0{,}276)^2 = 45 \cdot 0{,}0755 \cdot 0{,}0761 = 0{,}2585$$

$$\text{Prob}(x = 9) = \binom{10}{9} \cdot (0{,}724)^9 \cdot (0{,}276)^1 = 10 \cdot 0{,}0546 \cdot 0{,}276 = 0{,}1506$$

$$\text{Prob}(x = 10) = \binom{10}{10} \cdot (0{,}724)^{10} \cdot (0{,}276)^0 = 1 \cdot 0{,}0395 \cdot 1 = 0{,}0395$$

$$P_x = 0{,}4486$$

El proceso sería idéntico para el resto de las puntuaciones

$$p_c = \frac{\sum f_x (1 - 2(P_x - P_x^2))}{f_x} = \frac{21{,}3001}{25} = 0{,}852$$

$$p_a = 1 - 2\left[\frac{\sum f_x \cdot P_x}{N} - \left[\frac{\sum f_x \cdot P_x}{N}\right]^2\right] = 1 - 2\left[\frac{2{,}3929}{25} - \left[\frac{2{,}3929}{25}\right]^2\right] = 0{,}827$$

$$K = \frac{p_c - p_a}{1 - p_a} = \frac{0{,}852 - 0{,}827}{1 - 0{,}827} = \frac{0{,}025}{0{,}173} = 0{,}14$$

Puesto que el valor de Kappa es muy bajo, cabría esperar una fiabilidad baja.

5.

a)

Ítem	Juez 1	Juez 2	Juez 3
1	0,7	0,8	0,9
2	0,8	0,7	0,8
3	0,5	0,6	0,7
4	0,4	0,5	0,5
5	0,4	0,3	0,4

Los puntos de corte se calculan sumando las probabilidades, asignadas por cada uno de los jueces, de que cada uno de los ítems sea superado por los sujetos. Sumando dichas probabilidades tenemos:

Punto de Corte: 2,8 (Juez 1) 2,9 (Juez 2) 3,3 (Juez 3)

b)

El punto de corte del test es igual al promedio de los puntos de corte asignados por cada uno de los jueces.

$$P.C_{test} = \frac{2{,}8 + 2{,}9 + 3{,}3}{3} = \frac{9}{3} = 3$$

c)

El tercer juez, es el que considera el test más fácil ya que es el que define un punto de corte más alto. El primer juez, es el que considera el test más difícil ya que es el que define un punto de corte más bajo.

6.

a)

Ítem	Juez 1	Juez 2	Juez 3	Juez 4
1	*bcd*	*cd*	*bc*	*bcd*
2	*cd*	*bd*	*bcd*	*cb*
3	*ab*	*abd*	*abd*	*bd*
4	*acd*	*ac*	*cd*	*acd*

En primer lugar, debemos calcular la puntuación esperada por un sujeto en cada uno de los ítems del test. La puntuación esperada para un sujeto en un ítem viene dada como resultado de dividir la unidad por el número de alternativas del ítem que el sujeto no haya rechazado. A continuación sumamos las puntuaciones esperadas y su valor nos da el valor esperado en el test para cada juez. Estos datos son los que se recogen en la siguiente tabla:

Ítem	Juez 1	Juez 2	Juez 3	Juez 4
1	1/1	1/2	1/2	1/1
2	1/2	1/2	1/1	1/2
3	1/2	1/1	1/1	1/2
4	1/1	1/2	1/2	1/1
Σ	3	2,5	3	3

b)

El punto de corte del test es igual al promedio de los valores esperados para cada juez

$$P.C. = \frac{3+2,5+3+3}{4} = \frac{11,5}{4} = 2,87$$

$$P.C_{corregido} = A - \frac{N-A}{n-1} = 2,87 - \frac{4-2,87}{4-1} = 2,87 - 0,37 = 2,49$$

7. Soluciones a las preguntas conceptuales

1. Verdadera.
2. Falsa.

   Requiere una sola aplicación del test.
3. Verdadera.
4. Verdadera.
5. Verdadera.
6. Falsa.

   Tiene lugar cuando clasificamos incorrectamente a un sujeto dentro del grupo de no-maestría.
7. Verdadera.
8. Falsa.

   Proporciona una medida de la consistencia de clasificación de los sujetos independientemente del valor esperado por azar.
9. Verdadera.
10. Falsa.

    Se basa en que la probabilidad mínima de una decisión consistente es 0,50.
11. La afirmación es falsa.

    El punto de corte se establece en función del constructo objeto de estudio.
12. La afirmación es correcta

    No es necesario que los ítems sean de elección múltiple
13. La afirmación es falsa.

    Se trata de un método de compromiso.

9. BIBLIOGRAFÍA COMPLEMENTARIA

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

En el capítulo 21 se hace una exposición detallada de los tests referidos al criterio.

Muñiz, J. (1998). *Teoría clásica de los tests.* Madrid: Pirámide.

En el capítulo 2, el apartado 2.10 está dedicado al tema de la fiabilidad en los tests referidos al criterio.

TEMA 6

VALIDEZ DE LAS INFERENCIAS (I)

María Isabel Barbero García

SUMARIO

1. ORIENTACIONES DIDÁCTICAS

Cuando en el primer capítulo hablamos de la dificultad de medir variables psicológicas porque la gran mayoría de ellas no pueden ser observadas de forma directa y, por lo tanto, no se pueden medir directamente, dimos una solución al problema planteando que la medición se podía llevar a cabo por medio de indicadores. ¿Que queríamos decir con esto?, sencillamente que para poder medir este tipo de variables, a las que denominamos constructos teóricos, variables latentes o atributos psicológicos, entre otras acepciones, es necesario seleccionar una serie de conductas que representen algún aspecto de ese constructo y que sean consideradas indicadores del mismo. Estas conductas ya sí son observables de forma directa y, por lo tanto, pueden ser medidas mediante los instrumentos adecuados elaborados «ad hoc». Podremos decir que se ha obtenido una medida del constructo cuando se obtenga una medida de las conductas seleccionadas como indicadores.

Ahora bien, los instrumentos elaborados para medir estas conductas han de cumplir una serie de requisitos para que puedan ser utilizados con las suficientes garantías de calidad, entre estos requisitos hay dos fundamentales: que proporcionen medidas *fiables* a partir de las cuales se puedan hacer inferencias *válidas*.

La fiabilidad de las medidas, como se ha visto en el Tema 4, hace referencia al grado en que las puntuaciones obtenidas al aplicar los tests a un sujeto, o muestra de sujetos, reflejan su nivel real en el rasgo, o característica medida; es decir, al grado en que esas puntuaciones están libres de los *errores aleatorios* presentes en cualquier proceso de medición. La validez de las inferencias que se puedan hacer a partir de las puntuaciones obtenidas por los sujetos al aplicarles el test, problema que se abordará en éste y en el tema siguiente, se refiere al grado de relación que se puede establecer entre la evidencia empírica obtenida y el concepto teórico que se tiene del constructo que se intenta medir.

Alguna de las críticas que se han hecho en relación con la construcción y evaluación de los tests es que muchas veces el proceso ha estado orientado más a la obtención de instrumentos de me-

dición fiables que a la obtención de instrumentos válidos. A nuestro juicio, por muy fiables que sean las medidas que proporciona un test, si éstas no se refieren a aquello que se quiere medir difícilmente se podrán interpretar las puntuaciones obtenidas, de ahí la importancia del tema que vamos a estudiar.

En este tema, después de hacer una introducción al concepto de validez y a su evolución histórica se exponen las distintas clases de evidencia que se pueden obtener a la hora de llevar a cabo un proceso de validación: contenido, constructo y relativa al criterio, haciendo hincapié en cuál sería la más adecuada en cada caso y exponiendo los procedimientos estadísticos que van a permitir su obtención e interpretación.

Respecto a los estudios de validación referida al criterio, nos hemos centrado en la forma de llevarlos a cabo cuando hay un único predictor y una única variable criterio, dejando para el tema siguiente la forma de llevar a cabo el estudio de validación cuando se cuenta con varios predictores.

Es necesario que los alumnos aprendan a diferenciar claramente las distintas formas de llevar a cabo un estudio de validación y la forma de interpretar la evidencia obtenida; sólo así podrán estar en condiciones de interpretar las puntuaciones obtenidas por los sujetos en los tests y, a partir de ellas, tomar decisiones con una cierta garantía de éxito.

2. INTRODUCCIÓN AL CONCEPTO DE VALIDEZ Y SU EVOLUCIÓN HISTÓRICA

Al igual que el concepto de Psicometría ha evolucionado a lo largo de los años con la incorporación de los conocimientos científicos que han ido surgiendo a partir de las investigaciones realizadas; al concepto de validez, que por otra parte es un concepto psicométrico, le ha ocurrido lo mismo. Mientras que la medición de las características físicas como la longitud, por ejemplo, tal y como se viene haciendo, ha probado sus ventajas y utilidad y nadie plantea hoy día seriamente la necesidad de cambiar por otras formas de medición, no ocurre lo mismo con las características psicológicas (constructos teóricos) ya que, en ocasiones, la aparición de nuevos conocimientos puede aconsejar la modificación de la forma de medición de las mismas y la búsqueda de enfoques alternativos.

Aunque es difícil dar una definición concreta acerca del concepto de validez, convencionalmente y en relación con los tests, se acepta que el término hace referencia *al grado en que el test mide aquello que pretende medir*. En este sentido, un test será válido para medir razonamiento espacial, por ejemplo, si mide este tipo de razonamiento y no otra cosa. Ahora bien, cuando decimos que un test mide razonamiento espacial surgen una serie de interrogantes: ¿mide realmente

eso?, ¿en qué grado lo mide?, ¿mide sólo razonamiento espacial?, el intentar dar solución a estos interrogantes forma parte de los estudios de validación de los tests.

De la definición anterior se deduce que el concepto de validez hace referencia al grado de relación entre el test y el constructo que se quiere medir. En la medida en que la relación entre el test y el constructo que pretende medir sea más estrecha, el test será más válido. Ahora bien, es necesario aclarar, y lo haremos en más de una ocasión, que cuando hablamos de la relación entre el test y el constructo, en realidad estamos haciendo referencia a la relación entre las puntuaciones obtenidas por los sujetos en el test y la medida obtenida en el indicador o indicadores del constructo.

En esencia el concepto de validez no ha cambiado sustancialmente a lo largo de los años, lo que sí ha cambiado es la forma de abordar y operativizar esa relación entre el test y el constructo.

Hasta los años 50 del siglo pasado, los tests se valoraban fundamentalmente por su utilidad práctica, sobre todo para *la selección y clasificación* de personal. No se puede olvidar el éxito alcanzado con la utilización de los tests para la selección y clasificación de los reclutas en el Ejército de EE.UU. durante la Primera Guerra Mundial y su rápida implantación, a partir de entonces, en las empresas y escuelas de todo el país. Desde esta perspectiva la validez se entendía como la capacidad del test para *predecir* un criterio externo. Este criterio podía ser algún constructo teórico, como la aptitud para el vuelo, o el rendimiento futuro de los reclutas en un puesto de trabajo, por ejemplo en un puesto de mecánico. La forma de operativizar la relación entre el test y el criterio era mediante un coeficiente de correlación. Así, un test era válido en la medida en que existiera correlación entre las puntuaciones obtenidas por los sujetos en el test y las obtenidas en el criterio externo seleccionado. Al concepto de validez así entendido se le denominó *validez predictiva de los tests*.

EJEMPLO:

Supongamos que se desea llevar a cabo una selección de vendedores y, además de otras técnicas, se piensa utilizar un test en el proceso de selección. No se sabe si el test es válido o no, pero para que se pueda decir que el test tiene validez predictiva, deberá permitir diferenciar a los buenos de los malos vendedores distinguiendo los diferentes grados de *pericia o capacidad para las ventas (*constructo a medir). Para comprobar la validez predictiva del test es necesario seleccionar algún indicador (o indicadores) que permita obtener una medida del criterio externo; un indicador puede ser, por ejemplo, el número de ventas realizadas en una semana (variable observable relacionada con el constructo); una vez seleccionado el indicador, se aplicará el test a todos los aspirantes al puesto y, después de un cursillo sobre técnicas de ventas, se les pondrá a vender durante una semana; al cabo de la misma se les evaluará en función del número de ventas realizadas y ese dato será su medida en el criterio externo. Para comprobar si el test tiene validez predictiva se calculará la correlación entre las puntuaciones que han obtenido en el test todos los aspirantes y el

número de ventas realizadas en la semana de prueba; si la correlación es alta diremos que el test tiene validez predictiva, puesto que los que hayan obtenido puntuaciones altas en el test serán también los que hayan realizado un mayor número de ventas, y los que obtengan puntuaciones bajas en el test habrán realizado un número de ventas menor.

Ya se puede imaginar el lector la dificultad y el coste que supone el estudio de la validez predictiva de un test. En nuestro ejemplo supone aplicar el test a todos los aspirantes, darles un cursillo de formación en técnicas de ventas y tenerles a todos trabajando durante al menos una semana para poder tener una medida del criterio externo (el número de ventas). Esto a veces es imposible de llevar a cabo, o no tiene sentido hacerlo; por eso, poco a poco, fue surgiendo otra forma de estudiar la validez de los tests relacionada con criterios externos, *la validez concurrente*, que se diferencia de la validez predictiva en que la recogida de la información, tanto del test como del criterio, se hace simultáneamente.

EJEMPLO:

Vamos a seguir con el ejemplo anterior pero con un enfoque distinto. Queremos disponer de un test que sirva para hacer una selección de vendedores porque una empresa nos ha solicitado que hagamos una selección para cubrir cuatro puestos de trabajo. Entonces lo que se hace es lo siguiente: a una muestra de vendedores de las mismas características que los que exigen los puestos de trabajo a cubrir, se les aplica el test cuya capacidad predictiva se quiere estudiar y, al mismo tiempo, se pide a sus jefes directos que los evalúen en cuanto a su grado de pericia para las ventas (por ejemplo el número de ventas realizadas en la última semana). De esta manera las puntuaciones obtenidas por los vendedores en el criterio externo (número de ventas en la última semana) y las que han obtenido en el test se obtienen en el mismo momento temporal (validación concurrente). La correlación entre las dos series de puntuaciones, nos va a indicar si el test puede ser utilizado posteriormente para hacer la selección con ciertas garantías de éxito al avalar, en cierta medida, que los aspirantes que obtengan mejores resultados en el test serán buenos vendedores.

Hay veces que se puede obtener la medida del criterio con anterioridad a la del test; en este caso se habla de *validez retrospectiva*.

La forma de operativizar la relación entre el test y el criterio tanto en la validez predictiva como en la concurrente y en la retrospectiva es mediante un coeficiente de correlación, tal y como hemos visto.

Al mismo tiempo, y junto a las concepciones de la validez ligada a criterios externos se fue perfilando un nuevo enfoque de la validez relacionada con criterios internos al propio test: *la validez de contenido*. Esta nueva concepción surge porque hay muchos contextos en los que no interesa demasiado estudiar la utilidad de los tests para predecir otras variables y, por lo tanto, no tiene sentido la utilización de criterios externos. Esto ocurre, sobre todo, en los tests de conocimientos.

En este tipo de tests no se utilizan criterios externos con los que correlacionar las puntuaciones obtenidas, el planteamiento es distinto, se trata de estudiar hasta qué punto, a partir del contenido de los tests, se puede inferir el rendimiento en una determinada materia; el test en sí mismo constituye su propio criterio.

EJEMPLO:

Supongamos que se necesita preparar un test (un examen) para medir el conocimiento que los alumnos matriculados en la asignatura de Psicometría tienen de la materia. Esto que a primera vista puede parecer sencillo implica un esfuerzo por definir, en primer lugar, todos los contenidos propios de la Psicometría y, en segundo lugar, hacer un muestreo de cada uno de esos contenidos de manera que queden reflejados todos ellos en el test. Sólo de esta manera podremos tener cierta garantía de que el test tiene *validez de contenido*. No se podría preparar un test (examen) en el que sólo hubiera preguntas de fiabilidad, por ejemplo, ya que de las puntuaciones que obtuvieran los sujetos en el test no se podría inferir más que el grado de conocimientos de los sujetos acerca de la fiabilidad no de la Psicometría, puesto que el *dominio, universo* o *campo de contenidos* de la Psicometría es algo mucho más amplio.

> **Nota:** Cuando en el Tema 2 se abordó el problema de la construcción de instrumentos de medición psicológica ya se estudió la forma de elaborarlos de manera que tuvieran validez de contenido.

Tanto la validez predictiva como la concurrente dejaban muchos interrogantes sin responder, se sabía que el test, en nuestro ejemplo, valía para diferenciar realmente a los buenos de los malos vendedores, tenía utilidad práctica para llevar a cabo la selección, pero ¿por qué?, ¿qué es lo que realmente estaba midiendo el test?: ¿sería la capacidad de persuasión de los vendedores, su fluidez verbal, sus habilidades sociales, su extraversión, etc.? Ya la aparición de la validez de contenido marcó una nueva tendencia en los estudios de la validez al estar centrada más en qué es lo que mide el test que en su utilidad para predecir otras variables. Sin embargo, la respuesta real a todos esos interrogantes vendrá de la mano de otra nueva concepción de la validez, *la validez de constructo*. Este tipo de validez implica recoger toda la información necesaria para poder tener garantía suficiente de que las conductas observables que se han elegido como indicadores del constructo que se quiere medir, lo son realmente.

Todo esto nos hace reflexionar sobre la importancia que tiene, a la hora de construir un test, el definir claramente para qué se va a utilizar y qué es lo que se quiere medir; puesto que, en la medida en que el constructo esté mejor definido, será más fácil especificar qué conductas observables se van a utilizar como indicadores del mismo y, una vez especificadas estas conductas, se podrán tomar decisiones acerca de qué ítems (qué contenido) se van a incluir en el test para medirlas. Ahora bien, como señala Navas (2001), el que el constructo esté cuidadosamente definido facilita

las cosas, pero no nos exime de comprobar que, realmente, las puntuaciones obtenidas al aplicar el test miden esa característica o atributo y se pueden utilizar para el objetivo deseado, puesto que pueden estar midiendo además alguna característica no prevista e introduciendo un *error sistemático* en las puntuaciones obtenidas en el test.

EJEMPLO:

Supongamos que los ítems incluidos en el test utilizado en la selección de vendedores, además de medir las conductas relacionadas con la capacidad o pericia para las ventas, tienen una fuerte carga de rapidez y comprensión lectora; en este caso, los participantes en el proceso de selección que sean capaces de leer más deprisa, y tengan a su vez una mejor comprensión lectora, tendrán una mayor facilidad para contestar a los ítems que componen el test, con independencia de que sean mejores en el rasgo que éstos intentan medir.

El estudio de la validez de constructo del test permitirá responder a las preguntas que se habían planteado anteriormente: ¿mide el test aquello para lo que se construyó?, ¿mide sólo eso? También en este enfoque de la validez la forma de operativizar la relación entre el test y el constructo suele hacerse mediante técnicas correlacionales.

Estos cuatro tipos de validez: *predictiva, concurrente, de contenido y de constructo*, aparecen ya recogidos en el primero de una serie de documentos, publicado en 1954 por la American Psychological Association (APA): *Recomendaciones técnicas para los tests psicológicos y técnicas de diagnóstico (Technical Recommendations for Psychological Tests and Diagnostic Techniques)*, y elaborado por un comité de expertos con el objetivo de unificar, de alguna manera, los criterios que deben reunir los tests para poder ser utilizados como instrumentos científicos de medición. El presidente del comité fue Cronbach y uno de sus miembros Meehl que, en 1955, publicaron un artículo sobre la validez de constructo, en el que ya se empezaba a perfilar como el aspecto esencial de la validez que englobaría a todas las demás.

En el segundo documento publicado en 1955: *Recomendaciones técnicas para Tests de rendimiento (Technical Recommendations for achievement tests)*, intervinieron representantes de la American Educational Research Association (AERA) y el National Council on Measurement Used in Education (NCMUE) y fue publicado por la National Education Association (NEA).

El tercero, que vino a reemplazar a los dos anteriores, fue publicado por la APA en 1966 y preparado por un comité representante de la APA, AERA y el National Council on Measurement in Education (NCME) y se denominó: *Estándares para tests educativos y psicológicos y manuales (Standards for Educational and Psychological Tests and Manuals)*. En este documento, los cuatro tipos de validez quedaron reducidos a tres: validez de contenido, validez relativa al criterio y validez de constructo. En la validez referida al criterio quedaban subsumidas tanto la validez predictiva como la concurrente; también se asume que los distintos tipos de validez van unidos a objetivos concretos en el uso de los tests de ahí la importancia de definir cuales van a ser estos objetivos:

— Determinar el rendimiento o actuación de un sujeto en un universo de situaciones (contenido).

— Inferir el grado en el que un sujeto posee algún rasgo o atributo (constructo) que se supone vendrá reflejado por su ejecución en el test.

— Predecir el rendimiento o comportamiento futuro (predictiva) o estimar su rendimiento actual sobre una variable externa al test (concurrente).

La edición de 1974, cuyo título fue: *Estándares para Tests Educativos y Psicológicos (Standards for Educational and Psychological Tests*, AERA, APA y NCME*)*, supuso un avance en la definición del concepto de validez ya que, por primera vez, se afirma que *la validez se refiere a la adecuación de las inferencias que se realizan a partir de las puntuaciones de los tests u otras formas de medida*; se mantiene la distinción entre los tres tipos de validez y se consideran como formas independientes de interpretar las inferencias realizadas. Por otra parte se hace ya referencia explícita a que la validez no es una propiedad implícita a los tests ya que lo que se trata de validar no es el test en sí mismo sino las inferencias que se hagan a partir de las puntuaciones obtenidas por los sujetos.

En los *Estándares para la Evaluación Psicológica y Educativa (Standards for Educational and Psychological Testing, APA, AERA y NCME)* de 1985, y en los de 1999, ya se defiende una concepción unitaria de la validez, concepción que hace referencia *al grado en que la evidencia empírica obtenida y los conocimientos aportados por las teorías apoyan las inferencias que he hagan a partir de las puntuaciones obtenidas en el test cuando éste se utiliza para un objetivo concreto.*

Parece haber un acuerdo más o menos generalizado en que, desde el punto de vista científico, la única validez que se debe considerar es la validez de constructo y que las otras dos, la de contenido y la relativa al criterio, quedarían incluidas en ésta y serían consideradas estrategias de validación para comprender mejor lo que mide un test (Messick, 1989).

Ya no se habla de distintos tipos de validez, la validación de los tests es un proceso continuo que permite obtener distintos tipos de evidencia empírica, y un proceso de validación ideal debe incluir los tipos de evidencia implicados en los tres tipos tradicionales de validez: la de contenido, la de constructo y la relativa al criterio. Aunque siempre que se aplique un test psicológico es necesario llevar a cabo un estudio de validación de constructo (difícilmente se puede hacer ninguna inferencia si no se sabe lo que mide realmente el test), este tipo de validación no es siempre suficiente. Según sea la interpretación que se vaya a hacer de las puntuaciones obtenidas y el objetivo que se pretenda alcanzar al aplicar el test, será necesario obtener otros tipos de evidencia; así, por ejemplo, cuando se utilizan los tests en selección de personal, si el que una persona sea seleccionada depende de la predicción que se haga acerca de su rendimiento futuro en el trabajo, será necesario llevar a cabo un estudio de validación predictiva, y en los tests de conocimientos la estrategia fundamental sería la validación de contenido (Hambleton y Rogers, 1991).

La evolución del concepto de validez tuvo lugar gracias al esfuerzo de muchos autores, pero creo que es justo destacar algunos de los trabajos de Cronbach (1982, 1984, 1988) y Messick (1975, 1980, 1981, 1989) fundamentalmente.

Esta evolución en el concepto de la validez se puede observar también en las distintas ediciones del libro de Anastasi *Psychological Testing* (1954, 1961, 1968, 1976, 1982, 1988) y en las cuatro ediciones de *Essentials of Psychological Testing* (1949, 1960, 1970, 1984) de Cronbach.

Si consideramos que el término validez hace referencia a la adecuación de las inferencias realizadas a partir de las puntuaciones de los tests, es fácil definir la validación como:

> *el proceso mediante el cual el constructor, o el usuario de los tests, recoge la evidencia empírica necesaria para apoyar las inferencias que se van a realizar; entendiendo por evidencia tanto los datos, observaciones y hechos, como los argumentos que permitan apoyar y sustentar esos hechos.*

Si esto es así, para llevar a cabo un proceso de validación se requiere, en primer lugar, explicitar claramente el tipo de inferencia que se quiere realizar para, a continuación, diseñar el estudio empírico que permita obtener la información necesaria acerca del grado en que las puntuaciones obtenidas en el test (o los tests) son útiles para el tipo de inferencia requerida.

Siguiendo con las normas marcadas ya por los *Estándares* de 1985 y 1999, a lo largo del tema vamos a considerar la validez como un concepto unitario y el proceso de validación un proceso continuo que permitirá recoger la evidencia necesaria para poder interpretar las puntuaciones obtenidas al aplicar los tests para un determinado objetivo. En este sentido, no vamos a hablar de distintos tipos de validez, sino de distintas estrategias para obtener esa evidencia.

3. VALIDACIÓN DE CONTENIDO

Actualmente, para poder interpretar las puntuaciones de los tests la validez de contenido de los mismos es condición necesaria (Kane, 2009). Por otra parte, este tipo de validez hace referencia no sólo a los ítems que componen el instrumento de medida, sino a las instrucciones para su administración, corrección y puntuación (Abad, Olea, Ponsoda y García, 2011).

El objetivo que se persigue al llevar a cabo un estudio de validación de contenido es analizar hasta qué punto los elementos o ítems que componen el test son una muestra *relevante* y *representativa* del constructo sobre el que se van a realizar las inferencias.

En la definición anterior hemos destacado dos aspectos, *la relevancia y la representatividad* del constructo. El primero implica la necesidad de una clara y exhaustiva especificación de todas las posibles conductas observables que son representativas del contructo a medir (especificación del dominio de conductas); el segundo hace referencia a la necesidad de que todas esas conductas estén representadas en el test (representatividad del dominio).

EJEMPLO:

¿Qué quiere decir esto?

Vamos a representar el constructo que se quiere medir por una naranja y vamos a suponer que cada uno de los gajos de la naranja es una faceta o aspecto del mismo. Si quisiéramos construir un test para medir dicho constructo deberíamos hacer un análisis del tipo de conductas que podrían ser tomadas como indicadores de cada una de esas facetas (especificación del dominio de conductas) y, una vez seleccionadas todas esas conductas, deberíamos elegir una muestra representativa de ítems que permitieran medir cada una de ellas (representatividad del dominio).

Partiendo de esto, es fácil darse cuenta de que la distinción entre la validez de constructo y la de contenido es un poco artificial. En lo que se refiere a la especificación del dominio de conductas, o bien nos limitamos a describirlas simplemente, o en cuanto se intente establecer alguna definición operativa o formal entre esas conductas y el constructo se entra de lleno en el terreno de la validación de constructo. En lo referente a la representatividad del dominio, las investigaciones se han centrado, fundamentalmente, en los procedimientos de muestreo del dominio. Messick (1975) afirma que la especificación y representatividad del dominio son, en realidad, metas a conseguir a la hora de construir el test, pero que no son garantía de validez pues no proporcionan evidencia empírica para poder interpretar las puntuaciones.

Sin entrar en la polémica, diremos que cuando se lleva a cabo un estudio de validación del contenido de un test es necesario analizar hasta qué punto los elementos que lo componen son una muestra representativa de la clase de problemas o situaciones sobre las que se van a hacer inferencias y extraer conclusiones.

En el ámbito de la evaluación educativa, en los tests referidos al criterio (TRC) y en los denominados *tests de rendimiento académico,* las puntuaciones obtenidas se suelen utilizar para hacer inferencias acerca del grado en que los sujetos dominan un campo de conocimiento (dominio), no para hacer inferencias acerca de una conducta externa al test, ni acerca del rasgo o constructo medido. En estos tests, se pone de manifiesto el interés de los estudios de validación de contenido, ya que es relativamente fácil llevar a cabo la especificación del dominio (campo de conocimiento) sin hacer referencia al constructo. Las puntuaciones obtenidas se suelen utilizan para dar cuenta de si los sujetos han alcanzado un nivel mínimo de competencia en una determinada materia y la definición y especificación del dominio suele hacerse más en función de los objetivos instruccionales y educativos que se persigan que en referencia al constructo.

EJEMPLO:

Siguiendo con el ejemplo anterior, supongamos que nuestra naranja fuera la asignatura de Psicometría (campo de conocimiento). La especificación del dominio incluiría el análisis de todos aquellos componentes de la Psicometría que han de ser evaluados; por ejemplo, los alumnos deberán tener conocimientos de *fiabilidad, validez, análisis de ítems, interpretación de puntuaciones, etc.* Una vez especificado el dominio, para construir un test (examen) cuyo contenido sea válido, será necesario elaborar un conjunto de ítems que representen cada uno de esos componentes. El contenido del test será *relevante* si todos los ítems del test miden algún aspecto del dominio y no otra cosa, y será *representativo* si los ítems son una muestra representativa de todos los componentes especificados de la Psicometría; es decir, una muestra representativa del dominio.

La forma típica de llevar a cabo un estudio de validación de contenido, es utilizando un grupo de expertos que serán los encargados de analizar dos aspectos fundamentales:

— Que el test no incluya aspectos irrelevantes del dominio de interés.

— Que incluya todos los elementos importantes que definen el dominio (Livingston, 1977).

Se trata de hacer un análisis racional del contenido del test y, por lo tanto, los resultados del estudio estarán basados en los juicios subjetivos emitidos por los expertos.

Es necesario destacar la importancia que tiene la adecuada selección del grupo de expertos a la hora de establecer este tipo de validez; es necesario analizar las características y experiencia de los expertos en relación con el constructo tratado.

Para llevar a cabo la especificación del dominio, tal y como se ha expuesto en el Tema 2, es necesario, en primer lugar, analizar las áreas de contenido que se deben cubrir. En segundo lugar, se deben analizar los procesos que se van a evaluar y la importancia relativa de cada uno de ellos. En los tests educativos se deben establecer los objetivos instruccionales que se desean alcanzar. Una vez hecho esto, se puede construir una tabla de doble entrada en la que las columnas representen las distintas áreas de contenido (dominio) que definen el constructo a medir y las filas las distintas operaciones o procesos cognitivos implicados a la hora de responder a las preguntas o ítems de la prueba, o los distintos objetivos instruccionales. Las celdillas de esta tabla de doble entrada incluyen el porcentaje de ítems que debe contener la prueba en relación con cada área de contenido y cada proceso cognitivo empleado, u objetivo instruccional, para que se considere que el dominio está bien *representado* en el test.

Al margen del análisis cualitativo realizado por los expertos, para que el proceso de validación de contenido ofrezca información relevante es necesario que éstos aporten una valoración cuantitativa; por ello es necesario aplicar alguno de los métodos empíricos existentes que permitan cuantificar el grado de acuerdo que ha habido entre los expertos (Sireci, 1998).[1]

[1] Una revisión de los distintos procedimientos se puede encontrar en Pedrosa, Suárez-Álvarez y García-Cueto (2013).

Para evaluar la *relevancia* de los ítems en relación con el dominio se pueden utilizar varios procedimientos, uno de ellos es el propuesto por Hambleton (1980) que consiste en presentar a los expertos una serie de fichas cada una de las cuales contiene un ítem. Cada experto deberá expresar en una escala de 5 puntos el grado de ajuste de cada ítem con su correspondiente especificación en el dominio (conducta, área de conocimiento...), de manera que el 1 indique un mal ajuste y el 5 un ajuste muy bueno. Una vez hecho esto, se calcula la media o la mediana de los valores asignados por cada uno de los expertos del grupo a cada ítem, y el valor obtenido será el que indique el grado de relevancia del ítem. De esta forma se podrán seleccionar aquellos ítems que muestren un alto grado de ajuste y eliminar aquellos que por su bajo nivel de ajuste no sean relevantes.

La *representatividad* de los ítems que conforman el test hace referencia al grado en que se han cubierto las especificaciones del dominio, tanto en cuanto a los contenidos como a los objetivos propuestos. En la medida en que el dominio esté más y mejor representado, las inferencias que se puedan hacer acerca de la puntuación de los sujetos en el dominio, a partir de las puntuaciones que han obtenido en los tests, serán más precisas. Lo ideal sería poder contar con un banco de ítems referidos al dominio de interés y a partir del mismo extraer una muestra aleatoria de ítems; ahora bien, lo cierto es que no siempre es esto posible.

EJEMPLO:

Supongamos que un grupo de 100 expertos han de juzgar la relevancia de 3 ítems para medir la *calidad de la enseñanza* (constructo de interés). En la tabla adjunta se incluye la valoración asignada a cada uno de los ítems por el grupo de expertos:

ÍTEMS	ESCALA				
	1	2	3	4	5
A	0	10	10	60	20
B	20	40	30	10	0
C	10	20	50	10	10

Calcular la relevancia de cada ítem sabiendo que la categoría 1 indica un mal ajuste entre el ítem y el constructo y la categoría 5 un muy buen ajuste.

Calcularemos la mediana de cada ítem aplicando la siguiente fórmula:

$$Med. = L_i + I\frac{\left(NP/100\right) - f_b}{f_d}$$

Donde:

L_i = límite inferior del intervalo donde se encuentra la mediana.

I = amplitud del intervalo que en nuestro caso es la unidad.

$NP/100$ = 50% de la muestra.

f_d = número de sujetos de la muestra situados en el intervalo de la mediana.

f_b = número de sujetos de la muestra por debajo del intervalo de la mediana.

Para facilitar la comprensión del proceso se incluye la tabla de frecuencias acumuladas:

ÍTEMS	ESCALA				
	1	2	3	4	5
A	0	10	20	80	100
B	20	60	90	100	100
C	10	30	80	90	100

$$\text{Ítem } A = 3,5 + \frac{50-20}{60} = 4$$

$$\text{Ítem } B = 1,5 + \frac{50-20}{40} = 2,25$$

$$\text{Ítem } C = 2,5 + \frac{50-30}{50} = 2,90$$

Ante estos resultados se puede decir que el ítem A tiene un buen ajuste y, por lo tanto, se puede considerar como un ítem relevante para la medida de la calidad de la enseñanza, los otros dos no deberían incluirse puesto que el ajuste no es muy bueno.

4. VALIDACIÓN DE CONSTRUCTO

Este tipo de validación es, realmente, el que da significado a las puntuaciones de los tests, pues permite obtener evidencia de que las conductas observables que se han elegido como indicadores del constructo (variable latente inobservable) realmente lo son. Este tipo de estudios de validación permite responder, entre otras, a las siguientes preguntas: ¿mide realmente el test la variable que intenta medir? y ¿existe en realidad esa variable?

Partiendo de que los tests son instrumentos que permiten describir de forma indirecta, u operativizar, el grado en que los sujetos poseen alguna característica postulada a nivel teórico denominada *constructo*, la validación de constructo será el proceso que permitirá obtener evidencia acerca de la capacidad del test para medirle.

Este tipo de estudios de validación trata de garantizar científicamente que la variable que el test pretende medir es, efectivamente, una variable aceptable, cuyo concepto ofrece suficiente consistencia lógica dentro de un sistema teórico de la Psicología y descansa en suficientes comprobaciones experimentales que lo verifican (Yela, 1984).

Para llevar a cabo un estudio de validación del constructo es necesario:

- En primer lugar, definir cuidadosamente el constructo de interés a partir de las teorías que existan acerca del mismo, y postular una serie de hipótesis acerca de la naturaleza y grado de relación entre el constructo (variable latente inobservable) y una serie de variables (conductas directamente observables) y entre el constructo de interés y otros constructos.
- En segundo lugar, diseñar el instrumento de medida adecuado que habrá de contar con elementos relevantes y representativos de aquellas conductas que sean manifestaciones específicas y concretas del constructo.
- En tercer lugar, obtener datos empíricos de las relaciones entre las puntuaciones obtenidas al aplicar el test y las variables hipotetizadas (conductas observables).

Como se puede observar es necesario estudiar, por una parte, la relación entre el constructo y las conductas observables representativas del constructo; por otra, la relación entre el constructo y otros constructos y, finalmente, la relación entre esas conductas tomadas como indicadores del constructo y las puntuaciones obtenidas por los sujetos en el test.

Si se confirman las relaciones postuladas en las hipótesis planteadas, tal y como predice la teoría, se puede considerar que tanto el constructo como el test son útiles, en caso contrario será necesario hacer una nueva evaluación del constructo y/o de las demás variables incluidas en el estudio, o bien estudiar más detenidamente el marco teórico.

Los estudios de validación de constructo están centrados, fundamentalmente, en el análisis de la estructura del test, tanto interna como externa; es decir, en el estudio de las interrelaciones entre las puntuaciones obtenidas por los sujetos en los distintos ítems que conforman el test (estructura interna) y en las relaciones entre las puntuaciones obtenidas en el test y otras medidas del mismo constructo obtenidas en variables externas al mismo y consideradas relevantes (estructura externa).

Entre los métodos más utilizados para llevar a cabo la validación del constructo hemos de destacar el método de la matriz multimétodo-multirrasgo y el análisis factorial.

4.1. La matriz multimétodo-multirrasgo

Se trata de un método propuesto por Campbell y Fiske (1959) que permite evaluar la validez convergente y discriminante de los tests y analizar la estructura externa del test (o conjunto de tests).

Cuando un mismo constructo se mide por distintos tests la correlación (o correlaciones) entre las puntuaciones obtenidas nos dará una idea acerca de su *validez convergente*, en la matriz vendrá dada por los valores monorrasgo – multimétodo y se refiere, tal como se ha apuntado, al grado de relación entre los distintos tests que miden el mismo constructo. Cuando se miden distintos constructos con el mismo test, la correlación entre ellos debe ser baja, o al menos más baja que la correlación entre dos tests que midan el mismo constructo, este coeficiente de correlación será un indicador de su validez discriminante y en la matriz vendrá dada por los valores multirrasgo – monométodo.

La lógica del procedimiento propuesto por Campbell y Fiske es la siguiente: Se intenta medir un mismo constructo mediante distintos procedimientos y distintos constructos mediante el mismo procedimiento y, una vez obtenidas todas las medidas, calcular las intercorrelaciones entre ellas. Si las correlaciones entre las medidas obtenidas del mismo constructo a través de distintos procedimientos son altas, el constructo quedará validado y se dirá que existe *validez convergente*. Por otra parte, si estas correlaciones son significativamente más altas que las obtenidas al correlacionar las medidas de distintos constructos con el mismo procedimiento se dirá que existe *validez discriminante*.

Vamos a poner un ejemplo que clarifique el procedimiento propuesto por Campbell y Fiske.

EJEMPLO:

Supongamos que se quieren medir tres constructos: Razonamiento numérico (RN), Factor espacial (FE) y Razonamiento abstracto (RA) y se han elaborado una serie de pruebas con distinto formato: Verdadero-falso (V-F), Elección múltiple (E-M) y Frases incompletas (F-I) para medir cada uno de ellos. Tenemos, por lo tanto, tres constructos diferentes y tres procedimientos distintos para llevar a cabo la medición.

Para analizar la *validez convergente* y *discriminante*, se selecciona una muestra de sujetos a los que se aplican todas las pruebas, obteniéndose las puntuaciones de los mismos en cada constructo y mediante cada uno de los procedimientos; a partir de esas medidas se calculan todas las intercorrelaciones posibles, que pueden ordenarse de una forma similar a la matriz que se presenta a continuación y facilitará la explicación de los coeficientes de correlación obtenidos.

Los valores que se encuentran entre paréntesis en la diagonal de la matriz, representan los distintos coeficientes de fiabilidad. Se trata de la correlación entre las puntuaciones obtenidas al medir el mismo constructo mediante el mismo procedimiento (pueden ser dos tests paralelos, por ejemplo).

Los valores que aparecen en cursiva y subrayados son las correlaciones obtenidas al medir el mismo constructo por distintos procedimientos, la cuantía de estos valores ofrece información acerca de la validez convergente. Finalmente, los valores que aparecen en negrilla corresponden a las correlaciones obtenidas al medir distintos constructos con los mismos procedimientos. Para ver si existe evidencia de validez discriminante es necesario comparar los valores correspondientes a los índices de validez convergente (cursiva y subrayados) con los que aparecen en negrilla; dado que realmente los primeros son bastante más altos que los segundos podemos decir que, en efecto, hay evidencia de validez discriminante.

TABLA 5.1
Matriz multimétodo-multirrasgo

	V-F			E-M			F-I		
	RN	FE	RA	RN	FE	RA	RN	FE	RA
V-F									
RN	(.95)								
FE	**.20**	(.90)							
RA	**.30**	**.28**	(.92)						
E-M									
RN	.90	.31	.40	(.93)					
FE	.26	.87	.33	**.37**	(.94)				
RA	.43	.20	*.84*	**.26**	**.37**	(.88)			
F-I									
RN	.79	.27	.31	.77	.15	.23	(.89)		
FE	.11	.68	.22	.24	.67	.31	**.19**	(.93)	
RA	.19	.18	.50	.19	.33	.72	**.41**	**.30**	(.64)

Uno de los problemas que plantea el procedimiento de la matriz multirrasgo-multimétodo, es que no existe un criterio estadístico que permita tomar decisiones acerca de si un test tiene realmente validez convergente y discriminante, lo único que se puede decir es que parece haber evidencia de su existencia o de su ausencia. Actualmente, para poder obtener mayor información se está utilizando el análisis factorial confirmatorio.

4.2. El Análisis Factorial

Es quizás la técnica más utilizada, tanto en su vertiente exploratoria como confirmatoria, para poner a prueba las hipótesis planteadas acerca de la estructura interna del constructo y de las relaciones del mismo con otras variables. No vamos a hacer aquí una exposición exhaustiva de la técnica puesto que el tema rebasa los objetivos de este curso; sin embargo, sí queremos que nuestros alumnos entiendan su utilidad para el estudio de la validación de constructo.

Las medidas que proporcionan los tests pueden hacer referencia a variables unidimensionales o multidimensionales y, precisamente, el análisis factorial nos va a permitir descubrir la estructura que subyace a las puntuaciones obtenidas por los sujetos en los distintos ítems del test o en un conjunto de tests.

Cuando el análisis factorial se utiliza desde el enfoque exploratorio, no se establecen hipótesis previas acerca del número de dimensiones que subyacen al constructo, es la propia técnica la que nos aportará esta información. Desde el enfoque confirmatorio, se establecen *a priori* hipótesis acerca de la estructura subyacente y del número de dimensiones existentes, y mediante las técnicas oportunas se comprueba si se pueden aceptar las hipótesis propuestas.

Nota: Una exposición clara del análisis factorial puede encontrarse en Harman (1980), Ferrando (1993) y Martínez-Arias (1995) y Martínez Arias, Hernández y Hernández (2006).

Bajo el epígrafe Análisis Factorial (AF), se incluyen una serie de técnicas estadísticas que tienen por objetivo representar y explicar un conjunto de variables observables (ítems de un test, conjunto de tests, escalas, etc,) mediante un menor número de variables latentes o inobservables llamadas factores. Cada factor podría ser considerado como un constructo (variable latente) que vendría definido por las variables observables que lo conformaran, estas variables son las que van a permitir dar una interpretación psicológica al constructo (factor).

Para llevar a cabo un análisis factorial se parte de un conjunto de n medidas tomadas a la misma muestra de sujetos en un conjunto de variables observables (supongamos que son las puntuaciones obtenidas por una muestra de sujetos en los n ítems de un test) y, a partir de ellas, se obtiene una matriz ($n \times n$) con las intercorrelaciones entre todas ellas. Es a partir de esta matriz de correlaciones, cuando aplicando alguna de las técnicas estadísticas incluidas bajo el epígrafe de Análisis Factorial, se intenta identificar un número más reducido de variables latentes llamadas *factores*. Cuando en un mismo factor se agrupan múltiples indicadores del constructo, se obtiene evidencia de la validez convergente. Cuando en el análisis se han obtenido medidas de otros constructos y éstas aparecen agrupadas en distintos factores, se obtiene evidencia de la validez discriminante.

El ejemplo siguiente puede ayudarnos a comprender lo que queremos decir, se trata de un ejemplo ficticio y, por lo tanto, los resultados no son reales.

EJEMPLO:

Supongamos que a la matriz de correlaciones obtenida en el ejemplo anterior se la hubiera aplicado alguna de las técnicas incluidas bajo la denominación de Análisis Factorial y que la estructura factorial encontrada hubiera sido la siguiente:

Variables	Factor 1	Factor 2
RN (V-F)	.86	—
RN (E-M)	.75	—
RN (F-I)	.92	—
RE (V-F)	—	.82
RE (E-M)	—	.74
RE (F-I)	—	.63
RA (V-F)	.42	.33
RA (E-M)	.51	—
RA (F-I)	—	.54

¿Cómo se interpretan los resultados obtenidos?

Se puede observar que después de la factorización se han obtenido 2 factores. En el primero de ellos se agrupan las medidas correspondientes a las variables utilizadas como indicadores del constructo razonamiento numérico (RN) junto a dos correspondientes al constructo razonamiento abstracto (RA). El segundo factor está definido por todas las medidas correspondientes a las variables utilizadas como indicadores del constructo razonamiento espacial (RE) junto a otras dos correspondientes al razonamiento abstracto. Estos resultados parecen indicar que en realidad estamos ante dos constructos bien definidos; respecto al tercer constructo hipotetizado, sería necesario hacer una nueva evaluación del mismo, estudiar más detenidamente su marco teórico, o bien revisar los tests utilizados para su medición.

5. VALIDACIÓN REFERIDA AL CRITERIO

Este tipo de estudios de validación permiten obtener evidencia acerca del grado en que las puntuaciones obtenidas en el test pueden utilizarse eficazmente para hacer inferencias acerca del comportamiento real de los sujetos en un criterio que no puede ser medido directamente, bien por no estar disponible en el momento de la investigación, bien porque su medida pueda resultar difícil o costosa y, por lo tanto, sea aconsejable obtener información del mismo por otros procedimientos.

En los estudios de validación referida al criterio el objetivo principal es evaluar la hipótesis de relación entre test y criterio; la forma de analizar esta relación depende de muchos factores entre ellos la complejidad del criterio y la dificultad para definirle claramente. Para Crocker y Algina (1986) se suelen utilizar dos tipos de índices o medidas para describir la capacidad de un test o conjunto de tests para predecir un criterio: *medidas correlacionales* (coeficiente de validez, de determinación, de alienación, de valor predictivo, etc.) y *las medidas de error en la predicción* (errores de estimación).

Este tipo de estudios se suelen realizar desde dos perspectivas diferentes dependiendo del uso que se vaya a dar al test y del tipo de inferencias que se vayan a hacer. Cuando los tests se van a utilizar para la selección, clasificación o colocación de personas en determinados programas de formación o puestos de trabajo, lo interesante es analizar la *validez predictiva* de los tests; es decir, su capacidad para pronosticar, a partir de las puntuaciones obtenidas por los sujetos, su posterior rendimiento en el programa de formación, en el trabajo, en un curso de formación, etc. Si, por el contrario, se trata de utilizar los tests para hacer un diagnóstico, es más adecuado llevar a cabo un estudio de la *validez concurrente*.

Es necesario recordar que cuando se trata de obtener evidencia acerca de la validez predictiva de un test, la medida del criterio se obtiene con posterioridad a la del test; mientras que en los estudios acerca de la validez concurrente la medida del criterio se obtiene al mismo tiempo que la del test.

A diferencia de lo que ocurría en el proceso de validación de constructo, la validación referida al criterio es un proceso en el que la teoría no juega el papel principal, se acentúa el interés en el aspecto empírico del proceso más que en el teórico. No obstante, un análisis cuidadoso y una conceptualización teórica del criterio facilitan la tarea de aislar las dimensiones y subdimensiones que lo conforman, de manera que cada una de ellas pueda ser predicha por diferentes variables (validación de constructo del criterio). En otras palabras, como señalan Brogden y Taylor (1950), un estudio de validación de constructo del criterio ayudará a determinar las dimensiones a medir, cómo se medirá cada una de ellas y, si se desea, cómo combinarlas.

Para diseñar un estudio de validación referida al criterio es necesario seguir una serie de pasos:

1. Definir claramente el criterio que se quiere medir.
2. Identificar el indicador o indicadores que se van a utilizar para obtener la medida del criterio.
3. Seleccionar una muestra de sujetos que sea representativa de la población en la que posteriormente se va a utilizar el test.
4. Aplicar el test a la muestra de sujetos y obtener una puntuación para cada uno de ellos.
5. Obtener una medida de cada sujeto en el criterio bien en el mismo momento de la aplicación del test (validación concurrente) o bien al cabo de un cierto tiempo (validación predictiva).
6. Determinar el grado de relación entre las puntuaciones obtenidas por los sujetos en el test y la medida del criterio.

5.1. El problema de la selección y medición del criterio

Ya se ha comentado anteriormente que cuando los tests se utilizan para la selección, clasificación y colocación de las personas en determinados puestos de trabajo o programas específicos, los estudios de validación tienen como objetivo estudiar la efectividad con la que se puede pronosticar, a partir de las puntuaciones que hayan obtenido los sujetos en los tests, la eficiencia o éxito alcanzado en el puesto de trabajo o en el programa al que hayan sido admitidos. Se trata, por lo tanto, de utilizar los tests para seleccionar aquellas personas que vayan a tener una mayor probabilidad de realizar el trabajo, o aprovechar el programa con *éxito*.

Ahora bien, en este punto surge el problema de analizar qué es aquello que constituye el éxito. Este concepto es algo muy complejo (un constructo teórico) que tiene muchas facetas y, por lo tanto, es muy difícil de definir de forma precisa, y más difícil todavía obtener una medida adecuada y completa del mismo. Recordemos que en el ejemplo de la selección de vendedores, se utilizó como indicador del criterio de éxito el número de ventas realizadas en una semana, se trata de un indicador de tipo práctico, fácil de obtener, y probablemente de cara al cliente es un indicador válido. Supongamos ahora que hay que cubrir una plaza de profesor de Psicometría, en este caso sería más complejo determinar qué es lo que constituiría el éxito como profesor de Psicometría: ¿su conocimiento de la asignatura?, ¿su capacidad de empatía con los alumnos?, ¿la calidad de sus publicaciones?, ¿sus proyectos de investigación?, ¿su habilidad para la organización de las tareas propias de la asignatura?, etc. ,cada una de estas variables podrían ser consideradas indicadores del criterio de éxito o capacidad del profesor, pero son más difíciles de operativizar que el número de ventas en una semana. Ahora bien, tanto en un caso como en otro hay que tener en cuenta que *todos los indicadores son parciales* y no ofrecen una comprensión completa del criterio. Entonces, ¿cómo decidir cual es el indicador que se debe elegir?

Thorndike y Hagen (1989), consideran que los indicadores deben cumplir una serie de requisitos: *a)* que sean relevantes, *b*) que estén libres de sesgos, *c)* que sean fiables y *d*) que sean accesibles.

Se considera que un indicador es *relevante* en la medida en que se corresponde con el criterio. No hay evidencia empírica que nos permita decir si un indicador es relevante o no. Para apreciar la relevancia es necesario tener en cuenta consideraciones racionales y apoyarse en los juicios de expertos. La presencia de indicadores irrelevantes puede influir negativamente en las predicciones que se hagan y en las decisiones que se tomen. Por ejemplo: cuando un profesor está evaluando un examen de matemáticas de un niño, en el juicio que emita acerca de su capacidad pueden estar influyendo otros factores como la forma de presentación, o las faltas de ortografía. Estos factores pueden ser irrelevantes para la medida de la capacidad matemática del niño, y su influencia

puede atenuar la importancia del indicador seleccionado como relevante de aquello que se quiere predecir.

Un segundo requisito deseable es que los indicadores estén *libres de sesgos*; es decir, que las medidas del criterio representen la verdadera competencia de los sujetos y no estén determinadas por factores que actúen de manera diferencial en determinados grupos. Supongamos que se quiere evaluar la competencia de las secretarias de una empresa y se pide a sus jefes directos que las evalúen. El juicio de los jefes será un indicador libre de sesgos si la evaluación que hagan acerca de la competencia de sus secretarias no depende más que de su competencia profesional y no de «otros factores».

El tercer requisito es que sean *fiables,* las medidas del criterio que proporcionen los indicadores han de ser estables. Una medida de éxito en un determinado trabajo no puede variar de un día para otro. Una persona no puede ser considerada competente para el trabajo que realiza un día y al día siguiente ser considerado un incompetente. Si esto ocurriera; es decir, si la medida del criterio no fuera fiable, sería imposible encontrar un test capaz de pronosticarla.

Finalmente, los indicadores deben ser *accesibles.* A la hora de seleccionar los indicadores se suelen presentar problemas de distinta índole. Pueden ser problemas económicos, problemas debidos a que hay que esperar mucho tiempo para poder obtener la medida del criterio, etc. Todas estas limitaciones hay que tenerlas en cuenta a la hora de seleccionar los indicadores y tratar, en la medida de lo posible, de seleccionar aquellos que sean más accesibles siempre y cuando cumplan con los otros requisitos.

5.2. Procedimientos estadísticos utilizados en la validación referida al criterio

Para la exposición de este apartado nos basaremos en la realizada por Martínez – Arias (1995) y Martínez – Arias, Hernández y Hernández (2006). Cuando se quiere obtener un índice numérico que evidencie la validez de un test en relación con un criterio se pueden utilizar numerosos procedimientos, aunque los más utilizados están basados en correlaciones. No obstante, la utilización de una técnica u otra depende del diseño de recogida de datos para la validación y del número de variables implicadas: *a*) un único test predictor y un sólo indicador del criterio, *b*) varios predictores y un solo indicador del criterio, *c*) varios predictores cuantitativos y varios indicadores del criterio cuantitativos y *d*) procedimientos basados en la teoría de la decisión: validez y utilidad en las decisiones.

a) Un único test predictor y un solo indicador del criterio

Los procedimientos más utilizados son la correlación y el modelo de regresión lineal simple. Según sea la naturaleza de las variables implicadas se utilizará un tipo de correlación u otro (correlación de Pearson, biserial, biserial puntual, coeficiente phi, tetracórica, etc.).

b) Varios predictores y un sólo indicador del criterio

Hay veces que se utiliza una batería de tests para predecir un único criterio, en este caso los procedimientos que se utilizan son la correlación y la regresión lineal múltiple. Si el criterio es cualitativo, se suele utilizar otra técnica multivariante denominada *análisis discriminante* y cuando se utilizan criterios dicotómicos la *regresión logística.*

c) Varios predictores cuantitativos y varios indicadores del criterio cuantitativos

En este caso las técnicas más adecuadas son la regresión lineal multivariante y la correlación canónica. Sin embargo, rara vez se utilizan a la hora de llevar a cabo un estudio de validación debido a la dificultad para interpretar los resultados que proporcionan.

d) Procedimientos basados en la teoría de la decisión: validez y utilidad en las decisiones

Los procedimientos propuestos se basan en diferentes métodos para optimizar las decisiones realizadas con el test: técnicas maximin y minimax y especialmente la Teoría de la Utilidad Multiatributo.

No es posible la exposición de todas las técnicas por exceder a los objetivos de este libro. Expondremos aquellas que, a nuestro juicio, son las más importantes para que nuestros alumnos comprendan la forma de llevar a cabo un estudio de validación.

> **Nota:** El lector interesado en las técnicas de análisis multivariante puede consultar los siguientes textos en castellano: Bisquerra (1989), Cuadras (1981) y Sánchez-Carrión (1984). Una exposición introductoria a la Teoría de la decisión se puede encontrar en Macià, Barbero, Pérez-Llantada y Vila (1990).

6. VALIDACIÓN CON UN ÚNICO PREDICTOR Y UN SOLO INDICADOR DEL CRITERIO

Ya hemos comentado anteriormente que la correlación y la regresión lineal simple son, en este caso, las técnicas más utilizadas para obtener evidencia acerca del grado en que las puntuaciones obtenidas por los sujetos en el test pueden ser utilizadas para predecir las que obtendrían en el criterio. La correlación, porque nos permitirá conocer el grado de asociación entre el test y el criterio, y el modelo de regresión, porque nos permitirá pronosticar, a partir de las puntuaciones obtenidas en el predictor, las puntuaciones en el criterio.

Dado que nuestros alumnos ya han adquirido los conocimientos básicos acerca de este modelo, nosotros simplemente vamos a exponer su aplicación para estudiar las relaciones entre el test y el criterio.

6.1. El coeficiente de validez

Se define como la correlación entre las puntuaciones obtenidas por los sujetos en el test predictor y las obtenidas en el criterio. A partir de esta definición se pone de manifiesto la importancia que tiene el indicador elegido para obtener la medida del criterio ya que, en última instancia, a partir de las puntuaciones obtenidas por los sujetos en el test se podrán obtener tantos coeficientes de validez como indicadores del criterio se elijan para su validación, y un test puede ser muy válido para predecir un criterio cuando se utiliza un determinado indicador y obtener coeficientes de validez prácticamente nulos con respecto a otros.

El tipo de correlación utilizada para el cálculo del coeficiente de validez dependerá de la naturaleza de las variables implicadas, en el cuadro 6.1 se puede observar cuál es el índice más adecuado en cada caso.

CUADRO 6.1
Tipos de correlaciones en función del tipo de variables implicadas

INDICADOR CRITERIO	TEST		
	Continua	**Dicotomizada**	**Dicotómica**
Continua	Pearson	Biserial	Biserial puntual
Dicotomizada	Biserial	Tetracórica	$\phi_{biserial}$
Dicotómica	Biserial puntual	$\phi_{biserial}$	ϕ

Si designamos por *X* las puntuaciones del test y por *Y* las del indicador del criterio, la fórmula del coeficiente de validez será:

—*Correlación de Pearson:*

Cuando tanto el test (*X*) como el criterio (*Y*) son dos variables cuantitativas continuas:

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}} \qquad [6.2]$$

— *Correlación biserial:*

Siendo *X* la variable cuantitativa e *Y* la variable dicotomizada:

$$r_b = \frac{\bar{X}_p - \bar{X}_q}{S_x} \cdot \frac{pq}{y} = \frac{\bar{X}_p - \bar{X}}{S_x} \cdot \frac{p}{y} \qquad [6.3]$$

donde:

$\bar{X}_p$ = media en *X* de los que obtuvieron un 1 en *Y*.

$\bar{X}_q$ = media en *X* de los que obtuvieron un 0 en *Y*.

S_X = desviación típica en *X* de todas las personas de la muestra.

p y *q* = proporción de personas que obtuvieron un 1 y un 0 respectivamente en *Y*.

y = ordenada que en una distribución normal corresponde a la abscisa que divide el área total en dos partes iguales a «*p*» y «*q*».

$\bar{X}$ = media en *X* de todas las personas de la muestra.

— *Correlación biserial puntual:*

Siendo *X* la variable cuantitativa e *Y* la dicotómica:

$$r_{bp} = \frac{\bar{X}_p - \bar{X}_q}{S_x}\sqrt{pq} = \frac{\bar{X}_p - \bar{X}}{S_x}\sqrt{\frac{p}{q}} \qquad [6.4]$$

Los símbolos incluidos en esta fórmula tienen el mismo significado que los de la fórmula anterior.

— *Coeficiente ϕ:*

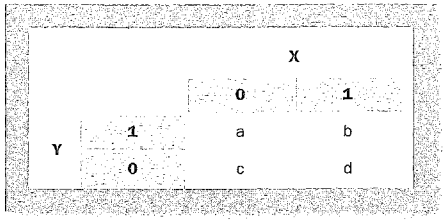
Las dos variables son dicotómicas

$$\phi = \frac{cb - ad}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \qquad [6.5]$$

donde:

a, *b*, *c* y *d* representan el número de personas de la muestra cuyas puntuaciones en *X* y en *Y* son respectivamente (0,1), (1,1), (0,0), (1,0).

Para una mejor comprensión se puede construir una tabla de doble entrada:



— *Coeficiente* $\phi_{biserial}$

La variable *X* se ha dicotomizado y la variable *Y* es dicotómica

$$\phi_{biserial} = \frac{bc - ad}{\sqrt{(a+b)(c+d)}} \cdot \frac{\sqrt{pq}}{y} \qquad [6.6]$$

donde:

a, *b*, *c* y *d* tienen el mismo significado que en la fórmula anterior y *p*, *q* e *y*, el mismo significado que en la correlación biserial. Se trata de un híbrido entre los dos coeficientes.

— *Correlación tetracórica*

Tanto la variable *X* como la *Y* son variables continuas que se han dicotomizado artificialmente.

El cálculo de la correlación tetracórica requiere la solución iterativa de una serie de potencias que incluye las potencias de r. Su cálculo es muy laborioso, pero se han ofrecido algunas aproximaciones muy sencillas, la más utilizada es calcular la razón *bc/ad* y consultar la tabla correspondiente que se ofrece al final del libro. Si la razón es menor que la unidad se debe usar la recíproca *ad/bc* para consultar la tabla, en este caso la correlación será negativa. El significado de *a*, *b*, *c* y *d* es el mismo que en las correlaciones anteriores (ver tabla de doble entrada anterior.)

Sea cual sea el coeficiente utilizado para calcular el coeficiente de validez, los valores que puede alcanzar van a estar incluidos en el intervalo –1 y 1.

6.2. El modelo de regresión lineal

Una vez conocido el grado de asociación entre el test y el criterio se puede utilizar el modelo de regresión para hacer pronósticos.

En los temas correspondientes al estudio de la fiabilidad se explicó la utilización del modelo de regresión lineal para hacer estimaciones acerca de la puntuación verdadera de los sujetos a partir de su puntuación empírica. Ahora vamos a ver de qué forma se va a utilizar el modelo para, a partir de las puntuaciones obtenidas por los sujetos en el test, hacer estimaciones acerca de su puntuación en el criterio.

Mediante el modelo de regresión se intenta buscar una ecuación lineal que haga mínimos los errores de pronóstico. Esta ecuación pondrá de manifiesto la relación de dependencia lineal entre el test y el criterio y tomará la siguiente forma:

$$y' = a + bX \qquad [6.7]$$

donde:

a = ordenada en el origen o término constante, que representa el valor pronosticado en el criterio (*Y'*) cuando en el test (*X*) se obtiene un valor cero.

b = pendiente de la recta de regresión, que representa el cambio en los valores del criterio *Y* por cada cambio unitario en el test *X*.

Nota: La exposición detallada del modelo la pueden encontrar nuestros alumnos en las unidades didácticas correspondientes a la asignatura de *Introducción al Análisis de Datos* y en las cle *Diseños de Investigación y Análisis de Datos*.

6.2.1. Ecuaciones de regresión

El valor de la pendiente se puede obtener en función del coeficiente de validez y de las desviaciones típicas de las puntuaciones obtenidas por los sujetos en el test y en el criterio:

$$b = r_{xy} \frac{S_Y}{S_X} \qquad [6.8]$$

Una vez calculado el valor de la pendiente se calcula el de la ordenada en el origen:

$$a = \overline{Y} - b\overline{X} \qquad [6.9]$$

La expresión anterior pone de manifiesto que la recta de regresión debe pasar por el punto ($\overline{X}$, $\overline{Y}$).

Una vez obtenidos los valores de *a* y *de b* se puede obtener la ecuación de la recta de regresión. Esta ecuación puede venir dada en tres tipos de puntuaciones: directas, diferenciales y típicas:

Ecuación en *puntuaciones directas*: $Y' = \left(\overline{Y} - r_{XY}\frac{S_Y}{S_X}\overline{X}\right) + r_{XY}\frac{S_Y}{S_X}X = r_{XY}\frac{S_Y}{S_X}\left(X - \overline{X}\right) + \overline{Y}$

Ecuación en *puntuaciones diferenciales*: $y' = r_{XY}\frac{S_Y}{S_X}\left(X - \overline{X}\right)$

Ecuación en *puntuaciones típicas*: $Z'_{Y'} = r_{XY}Z_X$

La diferencia entre la ecuación de regresión en puntuaciones directas y diferenciales es que en estas últimas la ordenada en el origen es cero y, por lo tanto, la ecuación pasa por el origen de coordenadas; al tener la misma pendiente se trata de dos rectas paralelas. Con respecto a la ecuación de regresión en puntuaciones típicas, al igual que la de puntuaciones diferenciales pasa por el origen de coordenadas y, teniendo en cuenta que en la escala de puntuaciones típicas la desviación típica es la unidad, la pendiente de la recta de regresión será el coeficiente de validez.

Hasta aquí, hemos utilizado una muestra de sujetos a la que se les ha aplicado el test cuya capacidad predictiva respecto al criterio se quiere validar; a estos mismos sujetos se les ha calificado en el criterio y, con todos esos datos, se han construido las ecuaciones de regresión. Pues bien, la verdadera utilidad de esas ecuaciones no está en pronosticar las puntuaciones de estos sujetos en el criterio, no tendría mucho sentido ya que conocemos las puntuaciones que realmente han obtenido, la verdadera utilidad está en la posibilidad de pronosticar la puntuación que obtendrán en el criterio otra muestra de sujetos, de las mismas características que la muestra utilizada para la construcción de las ecuaciones de regresión, a partir de las puntuaciones que obtengan en el test. Mediante la aplicación de las ecuaciones de regresión obtenemos una estimación puntual de las puntuaciones de los sujetos en el criterio.

6.2.2. La varianza residual o varianza error y el error típico de estimación

El coeficiente de validez indica la eficacia del test o variable predictora para estimar el criterio. En la medida en que el coeficiente de validez sea más alto, la estimación será más exacta; en el límite, cuando el coeficiente de validez fuera la unidad, el valor estimado coincidiría con la puntuación que realmente obtuvieran los sujetos en el criterio. Sin embargo, dado que nunca se alcanzan coeficientes de validez perfectos (iguales a 1 en valor absoluto), la estimación vendrá afectada por el denominado *error de estimación*. Así, se denomina error de estimación a la diferencia entre la puntuación que ha obtenido un sujeto en el criterio y la que se le pronostica mediante la ecuación de regresión $(Y - Y')$. Con cada sujeto se comete un determinado *error de estimación*. A la varianza de todos los errores de estimación cometidos con los sujetos de la muestra seleccionada se denomina *Varianza residual*, *Varianza error* o *Error cuadrático medio* y su fórmula es:

$$S^2_{Y \cdot X} = \frac{\sum (Y - Y')^2}{N} \qquad [6.10]$$

donde :

Y = puntuaciones obtenidas por cada sujeto en el criterio

Y' = puntuación pronosticada a cada sujeto mediante la ecuación de regresión

N = número de sujetos de la muestra

Esta varianza error representa la variabilidad media de las puntuaciones de los sujetos en el criterio respecto a la puntuación que se les pronostica mediante la recta de regresión. A la desviación típica de estos errores se denomina: *Error típico de estimación* y su fórmula es:

$$S_{Y \cdot X} = \sqrt{\frac{\sum (Y - Y')^2}{N}} \qquad [6.11]$$

Cuando se utilizan las ecuaciones de regresión para hacer los pronósticos se cumplen una serie de propiedades fundamentales:

- La media de las puntuaciones obtenidas por los sujetos en el criterio es igual a la media de las puntuaciones pronosticadas.
- La suma de todos los errores de estimación es cero, lo que implica que la media de los errores cometidos sea cero.
- La varianza de las puntuaciones obtenidas por los sujetos en el criterio (variable dependiente Y) es igual a la varianza de las puntuaciones pronosticadas más la varianza de los residuos o varianza error.

$$\overline{Y} = \overline{Y}'$$
$$\sum (Y - Y') = 0 \qquad [6.12]$$
$$S_Y^2 = S_{Y'}^2 + S_{Y \cdot X}^2$$

6.2.3. *Intervalos de confianza*

Debido a los errores de estimación que se cometen al hacer los pronósticos, más que estimaciones puntuales es conveniente hacerlas por intervalos; para ello, asumiendo que la distribución de dichos errores se ajusta a una distribución normal cuya desviación típica viene dada por el error típico de estimación, se establece un intervalo confidencial en torno a la puntuación pronosticada. Los pasos a seguir son los siguientes:

- Determinar un nivel de confianza y buscar su puntuación típica asociada.
- Calcular el error típico de estimación.
- Calcular el error máximo.
- Aplicar la ecuación de regresión correspondiente y obtener la puntuación pronosticada.
- Establecer el intervalo de confianza.

6.3. Interpretación de la evidencia obtenida acerca de la capacidad predictora del test

Ya hemos comentado que la varianza de las puntuaciones obtenidas por los sujetos en el criterio (varianza de la variable dependiente) se puede expresar como la suma de la varianza de las puntuaciones pronosticadas a partir de la variable predictora y la varianza de los residuos o varianza error.

$$\boxed{S_Y^2 = S_{Y'}^2 + S_{Y \cdot X}^2} \qquad [6.13]$$

A partir de esta ecuación se puede averiguar la proporción de la varianza de las puntuaciones de los sujetos en el criterio que se puede explicar a partir de la varianza de las puntuaciones en el predictor (varianza de las puntuaciones pronosticadas) y qué proporción no se puede explicar y corresponde a los residuos.

Si dividimos todos los términos de la ecuación por la varianza de las puntuaciones del criterio tendremos:

$$\boxed{1 = \frac{S_{Y'}^2}{S_Y^2} + \frac{S_{Y \cdot X}^2}{S_Y^2}} \qquad [6.14]$$

Ahora bien, en el segundo miembro de la ecuación, el primer término representa la proporción de la varianza del criterio que se puede pronosticar o predecir a partir del test o variable predictora y es igual al coeficiente de validez al cuadrado. Por lo tanto la expresión anterior se puede poner también como:

$$\boxed{\frac{S_{Y \cdot X}^2}{S_Y^2} = 1 - r_{XY}^2} \qquad [6.15]$$

y, a partir de ahí, deducir otra forma de expresión de la varianza error y del error típico de estimación:

$$\boxed{\begin{aligned} S_{Y \cdot X}^2 &= S_Y^2 \left(1 - r_{XY}^2\right) \\ S_{Y \cdot X} &= S_Y \sqrt{1 - r_{XY}^2} \end{aligned}} \qquad [6.16]$$

Cuando la escala que se utiliza es la de puntuaciones típicas, dado que la desviación típica es la unidad, la fórmula del error típico de estimación es:

$$\boxed{S_{Zx \cdot Zy} = \sqrt{1 - r_{XY}^2}} \qquad [6.17]$$

Una vez hecho este pequeño repaso, vamos a interpretar los resultados obtenidos en función de tres coeficientes:

6.3.1. *Coeficiente de determinación*

$$\boxed{C.D. = r_{XY}^2} \qquad [6.18]$$

Equivale al coeficiente de validez al cuadrado y representa la proporción (o el porcentaje) de la varianza de las puntuaciones de los sujetos en el criterio (variable dependiente) que se puede pronosticar a partir del test (variable predictora o independiente). También se define como la varianza común o asociada entre el test y el criterio.

6.3.2. Coeficiente de alienación

$$C.A. = K = \frac{S_{Y \cdot X}}{S_Y} = \sqrt{1 - r_{XY}^2} \qquad [6.19]$$

Aunque la fórmula es equivalente a la del error típico de estimación en puntuaciones típicas, de cara a la interpretación de este coeficiente conviene saber que, en realidad, indica la proporción que representa el error típico de estimación respecto a la desviación típica de las puntuaciones en el criterio. En la medida en que el error típico sea más pequeño que la desviación típica del criterio el coeficiente *K* será menor. El valor del coeficiente *K* oscila entre 0 y 1, será máximo cuando el coeficiente de validez sea 0 y será mínimo cuando el coeficiente de validez valga 1. El coeficiente de alienación al cuadrado es el complementario del coeficiente de determinación y representa, por lo tanto, la proporción (o el porcentaje si se multiplica por 100) de la varianza de las puntuaciones de los sujetos en el criterio que no se puede predecir a partir del test, es la proporción de varianza error que hay en la varianza de las puntuaciones de los sujetos en el criterio.

El coeficiente de alienación representa la inseguridad, o el azar, que afecta a los pronósticos.

6.3.3. Coeficiente de valor predictivo

$$C.V.P = 1 - \sqrt{1 - r_{XY}^2} \qquad [6.20]$$

Es el complementario del coeficiente de alienación y es otra forma de expresar la capacidad del test para pronosticar el criterio ya que representa la proporción (o el porcentaje si se multiplica por cien) de seguridad en los pronósticos.

6.3.4. Ejemplo

Supongamos que se quiere llevar a cabo un estudio de validación relativa al criterio de un test de aptitud mecánica *(X)*; para ello, se aplica a una muestra de sujetos representativa de la población en la que se va a utilizar el test. Estos sujetos son evaluados posteriormente por sus supervisores, en una escala de 0-10, utilizando como indicador de su capacidad mecánica el tiempo, medido en horas, que tarda cada uno en reparar un coche (*Y*) con la misma avería. Los resultados son los que aparecen en la tabla adjunta. (Téngase en cuenta que se trata de un ejemplo):

X	Y	X^2	Y^2	XY	Y'	(Y–Y')	$(Y–Y')^2$	
12	9	144	81	108	7,89	1,11	1,23	
14	7	196	49	98	8,68	–1,68	2,82	
15	10	225	100	150	9,08	0,92	0,85	
7	8	49	64	56	5,91	2,09	4,37	
9	5	81	25	45	6,71	–1,71	2,92	
4	4	16	16	16	4,73	–0,73	0,53	
61	43	711	335	473	43	0	12,72	Sumas

Tanto el test como la medida del criterio son variables cuantitativas, por lo tanto, para calcular el coeficiente de validez el índice más adecuado es la correlación producto-momento de Pearson.

— *El coeficiente de validez:*

$$r_{XY} = \frac{6 \cdot 473 - 43 \cdot 61}{\sqrt{\left[6 \cdot 711 - 61^2\right]\left[6 \cdot 335 - 43^2\right]}} = \frac{2.838 - 2.623}{\sqrt{545 \cdot 161}} = \frac{215}{296,22} = 0,73$$

Dado que el valor máximo del coeficiente de validez es la unidad, se puede deducir que el test tiene una buena capacidad predictiva. Más adelante se profundizará en la interpretación de los resultados obtenidos.

— *Las ecuaciones de regresión:*

Una vez obtenido el coeficiente de validez vamos a calcular las ecuaciones de regresión en puntuaciones directas, diferenciales y típicas teniendo en cuenta lo que se ha ido explicando anteriormente y los conocimientos que han de tener nuestros alumnos. Una vez construidas esas ecuaciones de regresión se pueden utilizar, posteriormente, para predecir las puntuaciones que obtendrán en el criterio otros sujetos, de las mismas características que los de la población sobre la que se construyeron, a partir de sus puntuaciones en el test. Para ello, basta sustituir el valor de *X* en la ecuación por las puntuaciones obtenidas por los sujetos. El resultado se recoge en la columna 6 de la tabla anterior. Comprobar que la media de las puntuaciones pronosticadas es igual que la de las puntuaciones obtenidas por los sujetos en el criterio.

En la columna 7 aparecen recogidos los errores de estimación cometidos con cada uno de los sujetos al hacer los pronósticos. Comprobar que la suma de estos errores es igual a cero. Hay que

recordar que si el coeficiente de validez hubiera sido la unidad, los errores de predicción o de estimación hubieran sido nulos.

Nota: Como ejercicio, los alumnos pueden calcular las puntuaciones pronosticadas en puntuaciones diferenciales y típicas.

Ecuaciones de regresión:

$$\overline{X} = \frac{\sum X}{N} = \frac{61}{6} = 10,17$$

$$\overline{Y} = \frac{\sum Y}{N} = \frac{43}{6} = 7,17$$

$$S_X^2 = \frac{\sum X^2}{N} - \overline{X}^2 = \frac{711}{6} - 10,17^2 = 118,5 - 103,43 = 15,07 \qquad S_X = 3,88$$

$$S_Y^2 = \frac{\sum Y^2}{N} - \overline{Y}^2 = \frac{335}{6} - 7,17^2 = 55,84 - 51,41 = 4,43 \qquad S_Y = 2,10$$

Puntuaciones directas: $Y = a + bX \Rightarrow Y = 3,15 + 0,395X$

$$b = r_{XY}\frac{S_Y}{S_X} = 0,73\frac{2,10}{3,88} = 0,395$$

$$a = \overline{Y} - b\overline{X} = 7,17 - 0,395 \cdot 10,17 = 3,15$$

Puntuaciones diferenciales: $y = bx \Rightarrow y = 0,395x$

Puntuaciones típicas: $Z_Y = r_{XY}Z_X = 0,73Z_X$

— *Error típico de estimación*

Vamos a comprobar cómo la varianza de las puntuaciones de los sujetos en el criterio es igual a la varianza de las puntuaciones pronosticadas más la varianza de los errores, para ello basta calcular las tres varianzas.

$$S_Y^2 = 4,43$$

$$S_{Y'}^2 = \frac{\sum Y'^2}{N} - \overline{Y'}^2 = \frac{322,36}{6} - 51,41 = 2,31$$

$$S_{Y \cdot X}^2 = \frac{\sum (Y - Y')^2}{N} - 0 = 2,12$$

$$S_Y^2 = S_{Y \cdot X}^2 + S_{Y'}^2 \Rightarrow 4,43 = 2,12 + 2,31$$

Hemos comentado que a la desviación típica de los errores de estimación se la denomina *error típico de estimación*, su valor en nuestro ejemplo será:

$$S_{Y \cdot X} = \sqrt{S_{Y \cdot X}^2} = \sqrt{2,12} = 1,46$$

— *Intervalos confidenciales*

Ya tenemos todos los datos para poder hacer una estimación acerca de la puntuación que se le pronosticaría a un sujeto en el criterio a partir de su puntuación en el test.

Supongamos que queremos saber qué puntuación le correspondería en el criterio a un sujeto que en el test hubiera obtenido una puntuación $X = 13$, y vamos a hacer una estimación puntual y por intervalos:

— Nivel de confianza 95% $\rightarrow Z_C = 1,96$

— El error típico de estimación ya estaba calculado: $S_{Y \cdot X} = 1,46$

— Error máximo = $Z_C \cdot S_{Y \cdot X} = 1,96 \cdot 1,46 = 2,86$

Para hacer el pronóstico en puntuaciones típicas hay que tener en cuenta que el error típico de estimación es distinto y hay que calcularlo.

$$S_{Zy \cdot Zx} = \sqrt{1 - r_{XY}^2} = \sqrt{1 - 0,73^2} = \sqrt{1 - 0,53} = 0,69$$

El error máximo en puntuaciones típicas será: 1,96 · 0,69 = 1,35

— Aplicación de las ecuaciones de regresión:

$Y = 3,15 + 0,395\ (13) = 8,28$ (en puntuaciones directas)

$y = 0,395\ (13 - 10,17) = 1,12$ (en puntuaciones diferenciales)

$Z_Y = 0,73 \cdot 0,73 = 0,53$ (en puntuaciones típicas)

$$Z_x = \frac{X - \bar{X}}{S_X} = \frac{13 - 10{,}17}{3{,}88} = 0{,}73$$

Las puntuaciones pronosticadas son la estimación puntual de las que obtendría en el criterio el sujeto que en el test obtuvo una puntuación de 13 puntos. Hacemos ahora la estimación por intervalos:

$$Y' \pm E_{máx.} = 8{,}28 \pm 2{,}86 \Rightarrow 5{,}42 \leq Y \leq 11{,}14$$
$$y' \pm E_{máx.} = 1{,}12 \pm 2{,}86 \Rightarrow -1{,}74 \leq y' \leq 3{,}98$$
$$Z_{Y'} \pm e_{máx.} = 0{,}53 \pm 1{,}35 \Rightarrow -0{,}82 \leq Z_Y \leq 1{,}88$$

A la vista de los resultados obtenidos podemos decir que la puntuación en el criterio de este sujeto estará comprendida en los intervalos encontrados y eso lo afirmamos con un nivel de confianza del 95%, o lo que es lo mismo con una probabilidad igual o menor de 0,05 de equivocarnos.

— *Coeficiente de determinación:*

$$C.D. = r_{XY}^2 = \frac{S_{Y'}^2}{S_Y^2} = \frac{2{,}31}{4{,}43} = 0{,}52$$

— *Coeficiente de alienación:*

$$C.A. = K = 0{,}69$$

— *Coeficiente de valor predictivo*:

$$C.V.P. = 0{,}31$$

Los valores obtenidos se pueden multiplicar por 100 y expresar como porcentajes. Así pues, hay un porcentaje de varianza común o asociada entre ambas variables de un 52%, o lo que es lo mismo, a partir de la variación de las puntuaciones obtenidas por los sujetos en el test se puede predecir el 52% de la variación de las puntuaciones de esos mismos sujetos en el criterio, quedando un 48% de la varianza de las puntuaciones en el criterio sin explicar por el test; es decir, un 48% de varianza error.

$$\frac{S_{Y.X}^2}{S_Y^2} = \frac{2{,}12}{4{,}43} = 0{,}48$$

Que el error típico de estimación representa el 69% de la desviación típica de las puntuaciones en el criterio, por lo tanto hay un porcentaje alto de inseguridad en los pronósticos frente al 31% de seguridad.

Nota: Las posibles diferencias encontradas pueden ser debidas a errores de redondeo.

7. EJERCICIOS DE AUTOEVALUACIÓN

1. Se quiere saber si un test de razonamiento abstracto tiene capacidad para predecir el rendimiento en matemáticas de los estudiantes de segundo de BUP. Para ello, se ha aplicado el test a una muestra de estudiantes obteniéndose una media y una desviación típica de 25 y 6 puntos respectivamente. Al finalizar el curso esos alumnos han sido evaluados por sus profesores en matemáticas obteniendo una media de 7 puntos y una varianza de 9.

Sabiendo que el 64% de la varianza de las puntuaciones de los sujetos en el criterio se puede predecir a partir del test. CALCULAR:

a) El coeficiente de validez del test.

b) Interpretar los resultados obtenidos en el punto anterior en función de los coeficientes de determinación, alienación y valor predictivo.

c) Varianza de los errores de estimación y error típico de estimación.

d) La varianza de las puntuaciones pronosticadas.

e) La puntuación directa, diferencial y típica que se le pronosticaría en matemáticas a un alumno que en el test hubiera obtenido una puntuación de 30 puntos.

f) Utilizando un nivel de confianza del 99%, establecer los intervalos confidenciales en torno a las puntuaciones obtenidas en el punto anterior.

2. Supongamos que se han intentado medir tres constructos diferentes a los que designaremos por A, B y C mediante tres métodos distintos y se quiere llevar a cabo un estudio de validación de constructo a través del análisis de la matriz multimétodo-multirrasgo. Supongamos que los resultados obtenidos al calcular las intercorrelaciones entre todas las puntuaciones obtenidas son los que se recogen en la matriz siguiente:

		MÉTODO 1			MÉTODO 2			MÉTODO 3		
		A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
	A_1	(.89)								
MÉT. 1	B_1	.49	(.93)							
	C_1	.35	.34	(.80)						
	A_2	**.60**	.21	.08	(.94)					
MÉT. 2	B_2	.22	**.65**	.11	.56	(.98)				
	C_2	.10	.12	**.49**	.58	.55	(.88)			
	A_3	.58	.22	.11	**.73**	.41	.34	(.99)		
MÉT. 3	B_3	.20	**.58**	.10	.40	**.68**	.23	.66	(.90)	
	C_3	.13	.13	**.61**	.36	.29	**.64**	.55	.59	(.95)

Comentar los resultados obtenidos.

3. Ejercicios conceptuales

Ante cada una de las afirmaciones que se muestran a continuación, el lector deberá responder si el concepto que contiene es verdadero o falso y justificar su respuesta.

1. El concepto de validez hace referencia a la estabilidad de las medidas obtenidas.
2. La validez es una propiedad intrínseca a los tests.
3. Un test puede tener varios coeficientes de validez.
4. La validez de constructo hace referencia al grado en que los elementos que componen el test miden realmente aquello que se quiere medir.
5. Para que un test tenga validez de contenido sus ítems deben ser relevantes y representativos del constructo que se quiere medir.
6. El error de estimación es la diferencia entre las puntuaciones obtenidas por los sujetos en el test y las obtenidas en el criterio.
7. El error típico de estimación es la varianza de los errores de estimación.
8. La validez de constructo representa la capacidad del test para pronosticar el criterio.
9. El coeficiente de validez puede ser negativo.
10. A medida que aumenta el coeficiente de determinación disminuye el coeficiente de alienación.

8. SOLUCIONES A LOS EJERCICIOS DE AUTOEVALUACIÓN

1.

a) A partir del enunciado del problema se puede obtener directamente el coeficiente de validez, ya que el porcentaje de varianza del criterio que se puede pronosticar a partir del test equivale al coeficiente de determinación que, a su vez, es el coeficiente de validez al cuadrado. Por lo tanto:

$$r_{XY}^2 = 0,64 \Rightarrow r_{XY} = \sqrt{0,64} = 0,80$$

b) A partir del coeficiente de validez obtenido podemos decir que:

— *El coeficiente de determinación es*: C.D. = 0,64

Indica que un 64% de la varianza de las puntuaciones de los sujetos en el criterio se puede pronosticar a partir del test; es decir, hay un 64% de varianza común o asociada entre el test y el criterio.

— *El coeficiente de alienación es*: C.A = $\sqrt{1-r_{XY}^2} = \sqrt{1-0,64} = 0,60$

Indica que en la desviación típica de las puntuaciones de los sujetos en el criterio, el 60% se debe a la desviación típica de los errores. Hay un 60% de inseguridad en los pronósticos. Elevado al cuadrado es el complementario del coeficiente de determinación e indica que hay un 36% de la varianza de las puntuaciones del criterio que no se puede pronosticar a partir del test. Se trata de la proporción (o porcentaje) de varianza error que hay en la varianza de las puntuaciones en el criterio.

— *El coeficiente de valor predictivo es:* C.V.P. = 1 – 0,60 = 0,40

Representa la proporción (o porcentaje) de seguridad en los pronósticos. Es el complementario del coeficiente de alienación. En nuestro ejemplo habrá un 40% de seguridad en los pronósticos.

c)

$$S_{Y \cdot X}^2 = S_Y^2\left(1-r_{XY}^2\right) = 9(1-0,64) = 3,24$$

$$S_{Y \cdot X} = \sqrt{S_{Y \cdot X}^2} = \sqrt{3,24} = 1,8$$

El error típico de estimación es la desviación típica de los errores de estimación. Se puede comprobar que si se divide por la desviación típica de las puntuaciones del criterio se obtiene el coeficiente de alienación.

d) Sabemos que la varianza de las puntuaciones de los sujetos en el criterio es igual a la varianza de las puntuaciones pronosticadas más la varianza de los errores. Por lo tanto:

$$S_{Y'}^2 = S_Y^2 - S_{Y \cdot X}^2 = 9 - 3,24 = 5,76$$

e) Para poder hacer cualquier pronóstico es necesario construir, en primer lugar, las correspondientes ecuaciones de regresión que tomarán la siguiente forma:

$Y' = a + bX$ (en puntuaciones directas)

$y' = bx$ (en puntuaciones diferenciales)

$Z_{Y'} = b^* Z_X$ (en puntuaciones típicas)

$$b = r_{XY}\frac{S_Y}{S_X} = 0,80\ \frac{3}{6} = 0,40$$

$$a = \overline{Y} - b\overline{X} = 7 - 0,40 \cdot 25 = -3$$

$$b^* = r_{XY} = 0,80$$

Una vez obtenidos los valores de las pendientes y de la ordenada en el origen se puede ya construir las ecuaciones de regresión:

$Y' = -3 + 0,40X$ (en puntuaciones directas)

$y' = 0,40\ x$ (en puntuaciones diferenciales)

$Z_{Y'} = 0,80\ Z_X$ (en puntuaciones típicas)

Nótese que la ecuación de regresión en puntuaciones diferenciales pasa por el origen de coordenadas y tiene la misma pendiente que la ecuación en puntuaciones directas. Respecto a la ecuación de regresión en puntuaciones típicas hay que decir que pasa por el origen de coordenadas y su pendiente es igual al coeficiente de validez.

Una vez construidas las ecuaciones de regresión sobre la muestra utilizada, se pueden aplicar para, a partir de las puntuaciones obtenidas en el test por una muestra de sujetos semejante a la anterior, hacer estimaciones de las que obtendrían en el criterio. En nues-

tro ejemplo queremos saber qué puntuación directa, diferencial y típica se le pronosticaría en el criterio a un sujeto que en el test hubiera obtenido 30 puntos. Basta sustituir los valores correspondientes en las ecuaciones de regresión:

$$Y' = -3 + 0,40(30) = 9 \qquad \text{(puntuación directa pronosticada)}$$

$$y' = 0,40(30 - 25) = 2 \qquad \text{(puntuación diferencial pronosticada)}$$

$$Z_{Y'} = 0,80\left(\frac{30 - 25}{6}\right) = 0,67 \qquad \text{(puntuación típica pronosticada)}$$

f) Al aplicar las ecuaciones de regresión se obtiene una estimación puntual de la puntuación de los sujetos en el criterio. Si se quiere precisar más es conveniente hacer una estimación por intervalos. Para ello, se escoge un determinado nivel de confianza, que en nuestro ejemplo es del 99%, y se procede como sigue:

$$\text{N.C. } 99\% \Rightarrow Z_C = \pm 2,58$$

— Se calcula el error típico de estimación:

$$S_{Y.X} = 1,8 \qquad \text{(en puntuaciones directas y diferenciales)}$$

$$S_{Zy.Zx} = \sqrt{1 - 0,64} = 0,60 \qquad \text{(en puntuaciones típicas)}$$

— Cálculo del error máximo en función del nivel de confianza:

$$E_{\text{máx.}} = Z_C \cdot S_{Y.X} = 2,58 \cdot 1,8 = 4,64 \qquad \text{(en puntuaciones directas y diferenciales)}$$

$$e_{\text{máx.}} = Z_C \cdot S_{Z_{y.Z_x}} = 2,58 \cdot 0,60 = 1,55 \qquad \text{(en puntuaciones típicas)}$$

— Intervalos confidenciales:

$9 \pm 4,64 \Rightarrow 4,36 \leq Y \leq 13,64$ (en puntuaciones directas)

$2 \pm 4,64 \Rightarrow -2,64 \leq y \leq 6,64$ (en puntuaciones diferenciales)

$0,67 \pm 1,55 \Rightarrow -0,88 \leq Z_Y \leq 2,22$ (en puntuaciones típicas)

Así se han obtenido los intervalos confidenciales dentro de los cuales se espera que se encuentren en el criterio las puntuaciones directa, diferencial y típica de un sujeto que en el test obtuvo una puntuación directa de 30 puntos. El intervalo se ha establecido con un nivel de confianza del 99%, o lo que es lo mismo, con una probabilidad igual o menor de 0,01 de error.

2. La matriz se puede analizar para ver si hay validez convergente y discriminante. La validez convergente vendrá dada por los coeficientes obtenidos al correlacionar las puntuaciones obtenidas al medir el mismo constructo con distintos métodos, a estos coeficientes se les denomina también *coeficientes monorrasgo-multimétodo*. En la matriz estos coeficientes aparecen en negrilla. Los valores que aparecen entre paréntesis en la diagonal de la matriz son los coeficientes de fiabilidad, también se les denomina *coeficientes monorrasgo-monométodo* porque se han obtenido al correlacionar las puntuaciones obtenidas al medir el mismo rasgo con el mismo método. Para ver si hay validez discriminante hay que analizar si los coeficientes de correlación obtenidos al medir el mismo rasgo con distintos métodos son mayores que los obtenidos al medir distintos rasgos con el mismo método. En la matriz estos coeficientes aparecen en cursiva y subrayados y reciben también el nombre de *coeficientes multirrasgo – monométodo*.

		MÉTODO 1			MÉTODO 2			MÉTODO 3		
		A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
MÉT. 1	A_1	(.89)								
	B_1	*.49*	(.93)							
	C_1	*.35*	*.34*	(.80)						
MÉT. 2	A_2	**.60**	.21	.08	(.94)					
	B_2	.22	**.65**	.11	*.56*	(.98)				
	C_2	.10	.12	**.49**	*.58*	*.55*	(.88)			
MÉT. 3	A_3	**.58**	.22	.11	**.73**	.41	.34	(.99)		
	B_3	.20	**.58**	.10	.40	**.68**	.23	*.66*	(.90)	
	C_3	.13	.13	**.61**	.36	.29	**.64**	*.55*	*.59*	(.95)

Se puede observar que hay validez convergente y discriminante, puesto que los coeficientes marcados en negrilla (monorrasgo-multimétodo) son altos y mayores que los marcados en cursiva y subrayados (multirrasgo-monométodo). Al resto de los coeficientes se les denomina multirrasgo-multimétodo.

3. Soluciones a los ejercicios conceptuales

1. La afirmación es falsa.

   Es el concepto de fiabilidad el que hace referencia a la estabilidad de las medidas. El concepto de validez hace referencia al grado en que el test mide aquello que pretende medir.

2. La afirmación es falsa.

   Durante muchos años se mantuvo la creencia de que la validez era una propiedad de los tests; sin embargo, hoy día se reconoce que el concepto de validez hace referencia a las inferencias que se hagan a partir de las puntuaciones obtenidas por los sujetos en los tests. De esta manera, un mismo test será válido para hacer determinadas inferencias y no tendrá ninguna validez para hacer otras.

3. La afirmación es verdadera.

   Dado que la validez no es una propiedad intrínseca a los tests, un test puede tener varios coeficientes de validez. Hemos definido el coeficiente de validez como la correlación entre las puntuaciones obtenidas por los sujetos en el test y las obtenidas en el indicador del criterio; pues bien, el coeficiente de validez dependerá, entre otros factores, del indicador seleccionado.

   En el tema siguiente se expondrán otros factores que afectan al coeficiente de validez.

4. La afirmación es verdadera.

   Cuando se lleva a cabo un estudio de validación de constructo se obtiene evidencia acerca de si realmente el test mide la variable que intenta medir, si mide sólo esa variable y si realmente esa variable tiene una consistencia a nivel teórico.

5. La afirmación es verdadera.

   La relevancia y la representatividad son dos aspectos que hay que tener en cuenta cuando se lleva a cabo un estudio de validación de contenido. La relevancia implica que los elementos que componen el test miden realmente algún aspecto o faceta del constructo (o área de conocimientos) que se quiere medir, y la representatividad implica que los ítems que conforman el test cubran todas las facetas especificadas del constructo (o área de conocimientos).

6. La afirmación es falsa.

   El error de estimación es la diferencia entre la puntuación empírica obtenida por los sujetos en el criterio y la que se les pronostica mediante la ecuación de regresión.

7. La afirmación es falsa.

   El error típico de estimación es la desviación típica de los errores de estimación.

8. La afirmación es falsa.

   Es la validez relativa al criterio la que permite obtener evidencia acerca de la capacidad de un test para predecir el criterio elegido.

9. La afirmación es verdadera.

   Dado que se trata de un coeficiente de correlación, los límites para el coeficiente de validez estarán entre –1 y 1.

10. La afirmación es verdadera.

    El coeficiente de alienación al cuadrado es el complementario del coeficiente de determinación, por lo tanto a medida que aumenta uno disminuye el otro.

9. BIBLIOGRAFÍA COMPLEMENTARIA

Hay bastantes textos a los que podrían acudir nuestros alumnos para el estudio de la validez, pero en castellano merecen destacar por orden alfabético los siguientes:

Martínez – Arias, M.R. (1995). *Psicometría: Teoría de los tests psicológicos y educativos.* Madrid: Síntesis. Capítulo 18.

Martínez – Arias, M.R.; Hernández Lloreda, M.J. y Hernández Lloreda, M.V. (2006). *Psicometría.* Madrid: Alianza Editorial. Capítulos 8 y 9.

Muñiz, J. (1998; 2002). *Teoría Clásica de los Tests.* Madrid: Pirámide. Capítulo 4.

Navas, M.J. (2001). *Métodos, diseños y técnicas de investigación psicológica.* Madrid: UNED. Capítulo 7.

Santisteban, C. (1990). *Psicometría.* Madrid: Norma. Capítulo 15.

TEMA 7

VALIDEZ DE LAS INFERENCIAS (II)

María Isabel Barbero García

SUMARIO

1. ORIENTACIONES DIDÁCTICAS

En el tema anterior se hizo una introducción al concepto de validez y a su evolución histórica para, a continuación, centrarnos en algunos de los procedimientos que se pueden utilizar para obtener la evidencia necesaria para hacer distintos tipos de inferencias. Se explicó de qué manera se podía llevar a cabo un estudio de validación de contenido y de constructo y, respecto a la forma de realizar un estudio de validación relativa al criterio, sólo se estudiaron los procedimientos estadísticos utilizados cuando se cuenta con un único predictor y un solo indicador del criterio. Esta situación es bastante rara cuando se trata de hacer una selección para un puesto de trabajo, o en otros muchos contextos aplicados; en estas situaciones lo normal es disponer de más de una variable predictora.

En este tema se estudiará la forma de llevar a cabo un estudio de validación cuando se utilizan varios predictores y también se expondrá la forma de analizar la validez de las decisiones tomadas a partir de las puntuaciones obtenidas por los sujetos en el test o en la batería de tests. Para finalizar el tema se expondrán algunos de los factores que afectan al coeficiente de validez y la forma de llevar a cabo un estudio de generalización de la validez.

Es importante que nuestros alumnos tengan muy claros y sepan interpretar los siguientes conceptos:

- Correlación múltiple.
- Correlación parcial y semiparcial.
- Error típico de estimación múltiple.
- Coeficientes de determinación, alienación y valor predictivo múltiples.
- Cómo construir y aplicar las ecuaciones de regresión múltiple en los distintos tipos de escalas (directa, diferencial y típica).
- Cómo seleccionar los predictores más adecuados de entre un conjunto de ellos.

- Cómo analizar la validez y utilidad de las decisiones tomadas a partir de las puntuaciones de los tests.
- Qué factores influyen en el coeficiente de validez y porqué.

2. VALIDACIÓN CON VARIOS PREDICTORES Y UN SOLO INDICADOR DEL CRITERIO

Si se quiere cubrir un puesto de trabajo en una empresa, un análisis serio de las necesidades y características del puesto de trabajo nos dará una idea de cuales son las aptitudes, conocimientos o variables de personalidad más adecuadas para desarrollar correctamente el trabajo exigido, y cuales impedirían el desarrollo correcto del mismo. Ahora bien, para llevar a cabo el análisis del puesto y conocer realmente cuales son las variables que van a incidir en que se desarrolle con éxito el trabajo, se puede hacer un estudio de validación que implica proceder de la siguiente manera:

Una vez seleccionadas, a priori, una serie de aptitudes, conocimientos o características de personalidad, por ejemplo, que son *aparentemente importantes* para el puesto de trabajo (variables predictoras), se seleccionan los instrumentos adecuados que van a permitir obtener una medida de cada una de ellas; esta medida obtenida de cada una de las variables predictoras se compara con la medida del criterio de éxito en el puesto de trabajo obtenida a partir de uno o varios indicadores. Es posible que las correlaciones entre las medidas obtenidas de las variables predictoras y la del criterio sean altas, que algunas no correlacionen con la medida del criterio y que, además, las medidas de las variables predictoras correlacionen entre sí. A partir de toda esta información se tendrá que decidir, en primer lugar, qué variables se consideran importantes y cuales se deben eliminar por no estar relacionadas con el criterio y, en segundo lugar, de qué forma se debe combinar la información obtenida a partir de las variables predictoras para que el pronóstico del éxito en el puesto de trabajo sea lo más efectivo posible.

Cuando se desea conocer el influjo de varias variables predictoras cuantitativas en otra también cuantitativa (criterio), los procedimientos estadísticos que van a permitir obtener esta información y dar solución a estos problemas son, fundamentalmente, la correlación múltiple y el modelo de regresión lineal múltiple.

> **Nota:** En el tema anterior ya se comentó que cuando las variables predictoras son cuantitativas y el criterio es discreto el procedimiento estadístico más adecuado es el análisis discriminante y si el criterio es dicotómico se podría utilizar la regresión logística. No vamos a entrar en la exposición de estos temas por exceder nuestros objetivos.

El modelo de regresión lineal múltiple permite obtener una ecuación de regresión, ponderando y combinando las variables predictoras seleccionadas, de manera que los errores de pronóstico que se cometan al estimar el criterio sean mínimos, y eliminando las variables que no aportan ninguna información relevante. Para poder hacer esto, es necesario introducir otros coeficientes de correlación como son: la correlación parcial y la correlación semiparcial que expondremos a continuación.

Vamos a hacer la introducción al tema utilizando sólo dos variables predictoras puesto que lo que nos interesa es que los alumnos comprendan la forma de proceder. La introducción de más variables predictoras complica mucho los cálculos y sería necesario utilizar notación matricial para resolver el problema y, desde luego, utilizar el software que hay para ello.

— Correlación parcial

Permite interpretar el grado de correlación entre la variable criterio (*Y*) y una de las variables predictoras, eliminando de antemano el efecto que *sobre dicha correlación* puedan estar ejerciendo el resto de las variables

$$R_{YX_1 \cdot X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{\left(1 - r^2_{YX_2}\right)\left(1 - r^2_{X_1X_2}\right)}} \qquad [7.1]$$

$$R_{YX_2 \cdot X_1} = \frac{r_{YX_2} - r_{YX_1} \cdot r_{X_1X_2}}{\sqrt{\left(1 - r^2_{YX_1}\right)\left(1 - r^2_{X_1X_2}\right)}}$$

En la primera se calcula la correlación entre la variable criterio *Y* y la variable predictora X_1 eliminando la influencia que, en esa correlación, pueda estar ejerciendo la variable X_2. En la segunda, al contrario, se calcula la correlación entre la variable criterio y la variable predictora X_2 eliminando el influjo que, en esa correlación, pueda estar ejerciendo la variable predictora X_1.

Si hubiera más de dos variables predictoras sería, por ejemplo: $R_{YX_1 \cdot X_2X_3X_4\ldots}$ Es decir, la correlación entre la variable criterio *Y* y la predictora X_1 eliminando del valor de esa correlación el efecto que puedan estar ejerciendo el resto de las variables predictoras.

— Correlación semiparcial

Permite conocer el grado de correlación entre la variable criterio (*Y*) y una de las variables predictoras, eliminando el efecto que *sobre esta variable predictora* puedan estar ejerciendo el resto de las variables

$$r_{Y(X_1 \cdot X_2)} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{1 - r^2_{X_1X_2}}} \qquad [7.2]$$

$$r_{Y(X_2 \cdot X_1)} = \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{\sqrt{1 - r^2_{X_1X_2}}}$$

La primera fórmula expresa la correlación entre la variable criterio (*Y*) y la variable predictora X_1 cuando de esta variable se elimina la influencia que pueda estar ejerciendo la variable predictora X_2. La segunda fórmula expresa la correlación entre la variable criterio (*Y*) y la variable predictora X_2 cuando de esta variable se elimina la influencia que pueda estar ejerciendo la variable predictora (X_1).

Si hubiera más variables predictoras la expresión sería, por ejemplo: $r_{Y(X_1 \cdot X_2 X_3 X_4 \ldots)}$ y representaría la correlación entre la variable criterio (*Y*) y la predictora X_1, después de haber eliminado de esta variable la posible influencia del resto de las variables predictoras.

Nota: No confundir la correlación parcial con la semiparcial. En la primera se elimina la influencia que, una variable predictora, ejerce sobre la correlación entre el criterio y la otra variable predictora. En la correlación semiparcial se elimina la influencia que una variable predictora ejerce sobre la otra variable predictora, no sobre la correlación.

2.1. El coeficiente de validez múltiple

Viene dado por la correlación múltiple entre las puntuaciones obtenidas por la muestra de sujetos en la variable criterio y las obtenidas en el conjunto de variables predictoras.

La correlación múltiple, va a permitir analizar el grado de asociación entre la variable dependiente (el criterio) y el conjunto de variables predictoras, en nuestro caso X_1 y X_2.

— *Correlación múltiple*

$$R_{Y \cdot X_1X_2} = \sqrt{\frac{r^2_{YX_1} + r^2_{YX_2} - 2r_{YX_1} r_{YX_2} r_{X_1X_2}}{1 - r^2_{X_1X_2}}} \qquad [7.3]$$

En nuestro caso, la correlación múltiple entre la variable criterio y el conjunto de las dos variables predictoras es igual a la raíz cuadrada de la suma de los cuadrados de las correlaciones simples entre el criterio y cada una de las variables predictoras menos el duplo del producto de las intercorrelaciones entre las tres variables, dividido por 1 menos la correlación al cuadrado entre las dos variables predictoras.

Otra forma de expresar la correlación múltiple es en función de los coeficientes de regresión múltiple en puntuaciones típicas, cuya fórmula expondremos más adelante y de las correlaciones de cada variable predictora con el criterio:

$$R_{Y \cdot X_1X_2} = \sqrt{b^*_1 r_{YX_1} + b^*_2 r_{YX_2}} \qquad [7.4]$$

donde:

Y = puntuaciones obtenidas por los sujetos de la muestra en el criterio.

X_1 y X_2 = puntuaciones obtenidas por los sujetos de la muestra en las dos variables predictoras.

b^*_1 y b^*_2 = coeficientes de regresión en puntuaciones típicas.

2.2. El modelo de regresión lineal múltiple

Aunque la estructura de este modelo es igual que la del modelo de regresión simple, las ecuaciones de regresión ya no son ecuaciones de una recta sino de un plano o hiperplano según que las variables predictoras sean dos o más.

Si tenemos *n* variables predictoras, la ecuación del hiperplano de regresión será:

$$Y' = a + b_1X_1 + b_2X_2 + \ldots\ldots\ldots b_nX_n$$

donde:

Y' = puntuación pronosticada en el criterio.

a = ordenada en el origen.

$b_1, b_2, \ldots b_n$ = coeficientes de regresión.

$X_1, X_2, \ldots X_n$ = variables predictoras.

Al igual que sucedía en el modelo de regresión lineal simple, para la construcción de las ecuaciones de regresión es necesario calcular los coeficientes. Cuando el número de variables predictoras es mayor que dos el problema se complica, por lo que se utilizan los programas de software adecuados.

Nota: Dado que el tema excede de los objetivos de nuestro curso no vamos a incluirla en este capítulo. El lector interesado puede consultar los textos en castellano de Martínez-Arias, 1995, Martínez – Arias y col. 2006 y Muñiz, 1998, 2002.

Nosotros vamos a poner un ejemplo muy sencillo para el caso en el que sólo se utilicen dos variables predictoras, ya que lo que nos interesa es que nuestros alumnos aprendan la lógica del procedimiento. Supongamos, por lo tanto, que se cuenta con las puntuaciones obtenidas por una muestra de sujetos en dos variables predictoras X_1 y X_2 y en un criterio Y. La ecuación del modelo de regresión, en este caso, sería:

$$Y = a + b_1X_1 + b_2X_2$$

donde:

a = ordenada en el origen del plano de regresión. Es el término independiente y equivale al valor que toma la variable criterio cuando $X_1 = X_2 = 0$.

b_1 = indica lo que aumenta el criterio al aumentar en una unidad la variable X_1 mientras permanece constante la variable X_2.

b_2 = indica el aumento del criterio cuando la variable X_2 aumenta en una unidad y permanece constante la variable X_1.

Los valores que deben alcanzar a, b_1 y b_2 deben ser aquellos que hagan mínimos los errores de pronóstico. Para su cálculo sería necesario resolver un sistema de ecuaciones o bien aplicar las fórmulas siguientes.

2.2.1. Ecuaciones de regresión

— *Puntuaciones típicas:*

$$Z_{Y'} = b_1^* Z_{X_1} + b_2^* Z_{X_2} \qquad [7.5]$$

$$b_1^* = \frac{r_{YX_1} - r_{YX_2} \cdot r_{X_1X_2}}{1 - r_{X_1X_2}^2} \qquad b_2^* = \frac{r_{YX_2} - r_{YX_1} \cdot r_{X_1X_2}}{1 - r_{X_1X_2}^2}$$

donde:

r_{YX_1}, r_{YX_2} = correlaciones entre la variable criterio (variable dependiente) y cada una de las variables predictoras (variables independientes).

$r_{X_1X_2}$ = correlación entre las dos variables predictoras.

$a = 0$

Al igual que ocurría en el modelo de regresión lineal simple, la ordenada en el origen es igual a cero, por lo tanto, se trata de un plano de regresión que pasa por el origen de coordenadas.

— *Puntuaciones diferenciales*

$$y' = b_1x_1 + b_2x_2 \qquad [7.6]$$

$$b_1 = b_1^* \frac{S_Y}{S_{X_1}} \qquad b_2 = b_2^* \frac{S_Y}{S_{X_2}} \qquad a = 0$$

La ordenada en el origen de la ecuación de regresión en puntuaciones diferenciales es igual a cero.

— *Puntuaciones directas*

$$Y' = a + b_1X_1 + b_2X_2 \qquad [7.7]$$

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$$

Los coeficientes b de regresión son iguales en puntuaciones directas y diferenciales.

Es importante remarcar dos cosas:

— las ecuaciones de los planos de regresión en puntuaciones directas y diferenciales corresponden a planos paralelos

— las ecuaciones de los planos de regresión en puntuaciones diferenciales y típicas pasan por el origen de coordenadas.

2.2.2. La varianza residual o varianza error y el error típico de estimación múltiple

El coeficiente de validez múltiple indica la eficacia de las variables predictoras para estimar el criterio. En la medida en que el coeficiente de validez sea más alto, la estimación será más exacta y, en el límite, cuando el coeficiente de validez fuera la unidad, el valor estimado coincidiría con la puntuación que realmente obtuvieran los sujetos en el criterio. Sin embargo, a pesar de haber utilizado más de una variable predictora, el coeficiente de validez no será perfecto (igual a 1 en valor absoluto) y la estimación vendrá afectada por el denominado *error de estimación* que equivale a la diferencia entre la puntuación que ha obtenido un sujeto en el criterio y la que se le pronostica mediante la ecuación de regresión $(Y - Y')$. Con cada sujeto se comete un determinado *error de estimación*. A la varianza de todos los errores de estimación cometidos con los sujetos de la muestra seleccionada se denomina *Varianza residual, Varianza error* o *Error cuadrático medio* y su fórmula es:

$$S^2_{Y \cdot X_1 X_2} = \frac{\sum (Y - Y')^2}{N} \qquad [7.8]$$

donde :

Y = puntuación obtenida por cada sujeto en el criterio.

Y' = puntuación pronosticada a cada sujeto mediante la ecuación de regresión múltiple.

N = número de sujetos de la muestra.

Esta varianza error representa la variabilidad media de las puntuaciones de los sujetos en el criterio respecto a la puntuación que se les pronostica mediante la recta de regresión. A la desviación típica de estos errores se denomina: *Error típico de estimación múltiple* y su fórmula es:

$$S_{Y \cdot X_1 X_2} = \sqrt{\frac{\sum (Y - Y')^2}{N}} \qquad [7.9]$$

2.2.3. Intervalos de confianza

Al igual que ocurría cuando sólo se utilizaba un predictor, más que estimaciones puntuales es conveniente hacerlas por intervalos debido a los errores de estimación que se cometen al hacer los pronósticos; para ello, asumiendo que la distribución de dichos errores se ajusta a una distribución normal cuya desviación típica viene dada por el error típico de estimación múltiple, se establece un intervalo confidencial en torno a la puntuación pronosticada. Los pasos a seguir son los siguientes:

- Determinar un nivel de confianza y buscar su puntuación típica asociada.
- Calcular el error típico de estimación múltiple.
- Calcular el error máximo.
- Aplicar la ecuación de regresión correspondiente y obtener la puntuación pronosticada.
- Establecer el intervalo de confianza.

2.3. Interpretación de la evidencia obtenida acerca de la capacidad predictora del conjunto de variables utilizadas

La varianza total de las puntuaciones obtenidas por los sujetos en el criterio (varianza de la variable dependiente) se puede expresar como la suma de la varianza de las puntuaciones pronosticadas a partir de las variables predictoras y la varianza de los residuos o varianza error. Dicho de otro modo, la variación total de los valores de Y es igual a la variación explicada por el influjo conjunto de X_1 y X_2 (el conjunto de variables predictoras) más la variación debida al azar o variación residual.

$$S^2_Y = S^2_{Y'} + S^2_{Y \cdot X_1 X_2} \qquad [7.10]$$

A partir de esta ecuación se puede averiguar la proporción de la varianza de las puntuaciones de los sujetos en el criterio que se puede explicar a partir de la varianza de las puntuaciones en las variables predictoras (varianza de las puntuaciones pronosticadas) y qué proporción no se puede explicar y corresponde a los residuos.

Si dividimos todos los términos de la ecuación (7.10) por la varianza de las puntuaciones del criterio tendremos:

$$1 = \frac{S^2_{Y'}}{S^2_Y} + \frac{S^2_{Y \cdot X_1 X_2}}{S^2_Y}$$

Ahora bien, la proporción de varianza del criterio que se puede explicar a partir de la variación debida al influjo conjunto de las dos variables predictoras es igual al coeficiente de validez al cuadrado. Entonces la expresión anterior se puede poner también como:

$$1 = R^2_{Y \cdot X_1 X_2} + \frac{S^2_{Y \cdot X_1 X_2}}{S^2_Y}$$

y, a partir de ahí, deducir otra forma de expresión de la varianza error y del error típico de estimación múltiple:

$$S^2_{Y \cdot X_1 X_2} = S^2_Y \left(1 - R^2_{Y \cdot X_1 X_2}\right)$$

$$S_{Y \cdot X_1 X_2} = S_Y \sqrt{1 - R^2_{Y \cdot X_1 X_2}} \qquad [7.11]$$

Cuando la escala que se utiliza es la de puntuaciones típicas, dado que la desviación típica es la unidad, la fórmula del error típico de estimación es:

$$S_{Z_Y \cdot Z_{X_1} Z_{X_2}} = \sqrt{1 - R^2_{Y \cdot X_1 X_2}} \qquad [7.12]$$

Vamos ahora a interpretar los resultados obtenidos en función de tres coeficientes:

2.3.1. Coeficiente de determinación múltiple

$$C.D. = R^2_{Y \cdot X_1 X_2} \qquad [7.13]$$

Equivale al coeficiente de validez múltiple al cuadrado y representa la proporción (o el porcentaje si se multiplica por cien) de la varianza de las puntuaciones de los sujetos en el criterio (variable dependiente) que se puede pronosticar a partir del conjunto de variables predictoras. También se define como la varianza común o asociada entre el criterio y las variables predictoras.

2.3.2. Coeficiente de alienación múltiple

$$C.A. = K = \frac{S_{Y \cdot X_1 X_2}}{S_Y} = \sqrt{1 - R^2_{Y \cdot X_1 X_2}} \qquad [7.14]$$

Aunque la fórmula es equivalente a la del error típico de estimación en puntuaciones típicas, de cara a la interpretación de este coeficiente conviene saber que, en realidad, indica la proporción que representa el error típico de estimación múltiple respecto a la desviación típica de las puntuaciones en el criterio. En la medida en que el error típico sea más pequeño que la desviación típica del criterio, el coeficiente *K* será menor. El valor del coeficiente *K* oscila entre 0 y 1, será máximo cuando el coeficiente de validez sea 0 y será mínimo cuando el coeficiente de validez valga 1. El coeficiente de alienación al cuadrado es el complementario del coeficiente de determinación y representa, por lo tanto, la proporción (o el porcentaje si se multiplica por cien) de la varianza de las puntuaciones de los sujetos en el criterio que no se puede predecir a partir del conjunto de variables predictoras, es la proporción de varianza error que hay en la varianza de las puntuaciones de los sujetos en el criterio. El coeficiente de alienación representa la inseguridad, o el azar, que afecta a los pronósticos.

2.3.3. Coeficiente de valor predictivo múltiple

$$C.V.P. = 1 - \sqrt{1 - R^2_{Y \cdot X_1 X_2}} \qquad [7.15]$$

Es el complementario del coeficiente de alienación y es otra forma de expresar la capacidad de las variables predictoras para pronosticar el criterio. Se interpreta como la proporción (o porcentaje) de seguridad con que se hacen los pronósticos.

2.3.4. Ejemplo

Se quiere averiguar si la fluidez verbal y la extraversión son dos variables que favorecen el número de ventas en un laboratorio farmacéutico. Para comprobarlo, se ha seleccionado una muestra de seis vendedores a los que se les han pasado dos pruebas, una de fluidez verbal (X_1) y otra de extraversión (X_2); asimismo, este grupo ha sido evaluado pos sus jefes en un criterio de pericia como vendedor, utilizando como indicador el número de ventas (en miles de euros) que realiza cada uno de ellos en un mes (Y).

Los resultados aparecen recogidos en las tres primeras columnas de la tabla 7.1:

TABLA 7.1
Datos del ejemplo

	Y	X_1	X_2	Y^2	X_1^2	X_2^2	YX_1	YX_2	X_1X_2
	4	2	4	16	4	16	8	16	8
	8	6	5	64	36	25	48	40	30
	6	5	6	36	25	36	30	36	30
	6	7	6	36	49	36	42	36	42
	5	4	7	25	16	49	20	35	28
	7	8	8	49	64	64	56	56	64
Suma	36	32	36	226	194	226	204	219	202

A partir de esos datos se calculan en primer lugar las intercorrelaciones entre las variables:

$$r_{Y\cdot X_1} = \frac{6\cdot 204 - 32\cdot 36}{\sqrt{\left[6\cdot 194 - 32^2\right]\left[6\cdot 226 - 36^2\right]}} = \frac{72}{91{,}65} = 0{,}79$$

$$r_{YX_2} = \frac{6\cdot 219 - 36\cdot 36}{\sqrt{\left[6\cdot 226 - 36^2\right]\left[6\cdot 226 - 36^2\right]}} = \frac{18}{60} = 0{,}30$$

$$r_{X_1X_2} = \frac{6\cdot 202 - 32\cdot 36}{\sqrt{\left[6\cdot 194 - 32^2\right]\left[6\cdot 226 - 36^2\right]}} = \frac{60}{91{,}65} = 0{,}65$$

A continuación calculamos:

— *Correlación múltiple*

$$R^2_{Y\cdot X_1X_2} = \frac{r^2_{YX_1} + r^2_{YX_2} - 2r_{YX_1}r_{YX_2}r_{X_1X_2}}{1 - r^2_{X_1X_2}} = \frac{0{,}79^2 + 0{,}30^2 - 2\cdot 0{,}79\cdot 0{,}30\cdot 0{,}65}{1 - 0{,}65^2} =$$

$$= \frac{0{,}406}{0{,}578} = 0{,}70$$

$$R_{Y\cdot X_1X_2} = \sqrt{0{,}70} = 0{,}84$$

— *Correlaciones parciales*

$$r_{YX_1\cdot X_2} = \frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{\sqrt{\left[1 - r^2_{YX_2}\right]\left[1 - r^2_{X_1X_2}\right]}} = \frac{0{,}79 - 0{,}30\cdot 0{,}65}{\sqrt{(1-0{,}30^2)(1-0{,}65^2)}} = \frac{0{,}595}{0{,}725} = 0{,}82$$

Este sería el valor de la correlación entre la variable criterio y la fluidez verbal habiendo eliminado de esa correlación el efecto de la variable extraversión. Antes de eliminar dicho efecto la correlación entre estas variables era de 0,79. Vemos por lo tanto, que el valor aumenta, lo que indica que la extraversión está influyendo negativamente.

Si calculamos la correlación entre la variable criterio y la extraversión eliminando de la correlación obtenida el efecto de la fluidez verbal, el valor obtenido será:

$$r_{YX_2\cdot X_1} = \frac{r_{YX_2} - r_{YX_1}r_{X_1X_2}}{\sqrt{\left[1 - r^2_{YX_1}\right]\left[1 - r^2_{X_1X_2}\right]}} = \frac{0{,}30 - 0{,}79\cdot 0{,}65}{\sqrt{\left(1-0{,}79^2\right)(1-0{,}65^2)}} = -0{,}46$$

La correlación obtenida es negativa, a diferencia del valor encontrado antes de eliminar la influencia de la fluidez verbal. Esto indica que la fluidez verbal estaba influyendo positivamente en la correlación.

— *Correlaciones semiparciales*

$$r_{Y(X_1\cdot X_2)} = \frac{r_{Y\cdot X_1} - r_{YX_2}\cdot r_{X_1X_2}}{\sqrt{1 - r^2_{X_1X_2}}} = \frac{0{,}79 - 0{,}30\cdot 0{,}65}{\sqrt{1 - 0{,}65^2}} = \frac{0{,}595}{0{,}759} = 0{,}78$$

$$r_{Y(X_2\cdot X_1)} = \frac{r_{YX_2} - r_{YX_1}r_{X_1X_2}}{\sqrt{1 - r^2_{X_1X_2}}} = \frac{0{,}30 - 0{,}79\cdot 0{,}65}{\sqrt{1 - 0{,}65^2}} = \frac{-0{,}214}{0{,}759} = -0{,}28$$

Cuando se elimina el influjo que una variable predictora tiene sobre la otra, la correlación obtenida varía ostensiblemente. Hay que recordar que la correlación entre las dos variables predictoras es bastante alta (r = 0,65). En lo posible, hay que evitar que la correlación entre las dos variables predictoras sea alta, de esta manera se podrá explicar un mayor porcentaje de varianza del criterio. Más adelante explicaremos el problema.

— *Ecuaciones de regresión en puntuaciones típicas:*

$$Z'_Y = b_1^* Z_{X_1} + b_2^* Z_{X_2}$$

$$a = 0$$

$$b_1^* = \frac{0,79 - 0,30 \cdot 0,65}{1 - 0,65^2} = \frac{0,59}{0,58} = 1,02$$

$$b_2^* = \frac{0,30 - 0,79 \cdot 0,65}{1 - 0,65^2} = \frac{-0,21}{0,58} = -0,36$$

$$Z'_Y = 1,02\ Z_{X_1} - 0,36\ Z_{X_2}$$

— *Ecuaciones de regresión en puntuaciones diferenciales*

$$b_1 = b_1^* \frac{S_Y}{S_{X_1}} \qquad b_2 = b_2^* \frac{S_Y}{S_{X_2}} \qquad a = 0$$

$$S_Y = \sqrt{\frac{\sum Y^2}{N} - \left(\frac{\sum Y}{N}\right)^2} = \sqrt{\frac{226}{6} - \left(\frac{36}{6}\right)^2} = 1,30$$

$$S_{X_1} = \sqrt{\frac{\sum X_1^2}{N} - \left(\frac{\sum X_1}{N}\right)^2} = \sqrt{\frac{194}{6} - \left(\frac{32}{6}\right)^2} = 1,98$$

$$S_{X_2} = \sqrt{\frac{\sum X_2^2}{N} - \left(\frac{\sum X_2}{N}\right)^2} = \sqrt{\frac{226}{6} - \left(\frac{36}{6}\right)^2} = 1,30$$

$$b_1 = 1,02\ \frac{1,30}{1,98} = 0,66 \qquad b_2 = -0,36\ \frac{1,30}{1,30} = -0,36 \qquad a = 0$$

$$y' = 0,66\ x_1 - 0,36\ x_2$$

— *Ecuación de regresión en puntuaciones directas*

$$Y' = a + b_1 X_1 + b_2 X_2$$

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$$

$$\overline{Y} = \frac{\sum Y}{N} = \frac{36}{6} = 6 \qquad \overline{X}_1 = \frac{\sum X_1}{N} = \frac{32}{6} = 5,33 \qquad \overline{X}_2 = \frac{\sum X_2}{N} = \frac{36}{6} = 6$$

$$a = 6 - 0,66 \cdot 5,33 + 0,36 \cdot 6 = 4,64$$

$$Y' = 4,64 + 0,66\ X_1 - 0,36\ X_2$$

— *Varianza error o varianza residual*

Hemos visto que hay dos formas de calcularla, bien a partir de las diferencias cuadráticas medias entre las puntuaciones obtenidas en el criterio y las pronosticadas mediante la ecuación de regresión múltiple, o aplicando la fórmula siguiente:

$$S^2_{Y \cdot X_1 X_2} = S^2_Y \left(1 - R^2_{Y \cdot X_1 X_2}\right) = 1,67\,(1 - 0,70) = 0,50$$

$$S^2_y = \frac{226}{6} - 36 = 1,67$$

Dado que se trata de un ejemplo vamos a hacerlo de las dos formas para comprobar que el resultado es el mismo. Por eso, en primer lugar, es necesario pronosticar las puntuaciones mediante la ecuación de regresión construida, sustituyendo en la misma los valores que han obtenido los sujetos en las dos variables predictoras. Por ejemplo, para el primer sujeto sería:

$$Y' = 4,64 + 0,66(2) - 0,36(4) = 4,52$$

A continuación se incluye la tabla 7.2 con los datos necesarios y, una vez aplicada la fórmula correspondiente, se puede observar que los resultados coinciden. Señalar también cómo la suma de las puntuaciones pronosticadas es igual que la de las puntuaciones empíricas obtenidas y cómo la suma de los errores de estimación o de pronóstico es cero.

TABLA 7-2
Datos necesarios para los análisis

	Y	X_1	X_2	Y'	(Y–Y')	(Y–Y')²
	4	2	4	4,52	–0,52	0,27
	8	6	5	6,8	1,2	1,44
	6	5	6	5,78	0,22	0,05
	6	7	6	7,1	–1,1	1,21
	5	4	7	4,76	0,24	0,06
	7	8	8	7,04	–0,04	0,00
Suma	36	32	36	36	0,00	3,03

$$S^2_{Y\cdot X_1X_2} = \frac{\sum (Y - Y')^2}{N} = \frac{3,03}{6} = 0,51$$

— *Error típico de estimación múltiple*

$$S_{Y\cdot X_1X_2} = S_Y\sqrt{1 - R^2_{Y\cdot X_1X_2}} = 1,30\sqrt{1 - 0,70} = 0,71$$

En la fórmula del error típico en puntuaciones típicas la desviación típica del criterio es la unidad, por lo tanto el valor de este error será:

$$S_{Z_Y\cdot Z_{X1}Z_{X2}} = \sqrt{1 - 0,70} = 0,548 \approx 0,55$$

— *Intervalos confidenciales*

Una vez construidas las ecuaciones de regresión, y conocido el error típico de estimación, es posible utilizarlas para pronosticar la puntuación que obtendrán en el criterio de pericia de ventas un grupo de sujetos pertenecientes a la misma población de la que se extrajo la muestra que sirvió para su construcción, a partir de las puntuaciones que obtengan en las variables predictoras.

Con los datos que hemos ido obteniendo vamos a calcular la puntuación que se le pronosticaría en el criterio a una persona que hubiera obtenido en la prueba de fluidez verbal 9 puntos y en la de extraversión 6. Para establecer los intervalos confidenciales utilizaremos un nivel de confianza del 99%.

Los pasos a seguir son:

NC 99% $\Rightarrow Z_C = 2,58$

$S_{Y\cdot X_1X_2} = 0,71$ (En puntuaciones típicas = 0,55)

Error máximo = $Z_C \cdot S_{Y\cdot X_1X_2} = 2,58 \cdot 0,71 = 1,83$ (En puntuaciones típicas = 1,42)

Aplicando las ecuaciones de regresión múltiple obtendremos la puntuación pronosticada al sujeto:

Puntuación típica:

$$Z_{Y'} = 1,02\ Z_{X_1} - 0,36\ Z_{X_2} = 1,02\left(\frac{9 - 5,33}{1,98}\right) - 0,36\left(\frac{6 - 6}{1,30}\right) = 1,89$$

Puntuación diferencial:

$$y' = 0,66\ x_1 - 0,36\ x_2 = 0,66 \cdot (9 - 5,33) - 0,36\ (6 - 6) = 2,42$$

Puntuación directa:

$$Y' = 4,64 + 0,66\ X_1 - 0,36\ X_2 = 4,64 + 0,66 \cdot 9 - 0,36 \cdot 6 = 8,42$$

Esta sería una estimación puntual pero vamos a hacer una estimación por intervalos, para ello a la puntuación pronosticada le sumamos y restamos el error máximo:

En puntuaciones típicas:

$$1,89 \pm 1,42 \Rightarrow 0,47 \leq Z_Y \leq 3,31$$

En puntuaciones diferenciales:

$$2,42 \pm 1,83 \Rightarrow 0,59 \leq y \leq 4,25$$

En puntuaciones directas:

$$8,42 \pm 1,83 \Rightarrow 6,59 \leq Y \leq 10,25$$

Estos intervalos marcan los límites entre los que se encontrará la puntuación en el criterio del sujeto seleccionado, con una probabilidad de acierto del 99% o, lo que es lo mismo, con una probabilidad igual o menor de 0,01 de equivocarnos.

— *Interpretación de los resultados obtenidos*

El coeficiente de determinación múltiple, viene dado por la correlación múltiple al cuadrado y es igual a 0,70. Esto indica que, a partir de las puntuaciones obtenidas por los sujetos en las dos variables utilizadas como predictores se puede explicar el 70% de la variación de las puntuaciones de los sujetos en el criterio, o lo que es lo mismo, que entre el criterio y el conjunto de variables predictoras hay un 70% de varianza común o asociada

El coeficiente de alienación múltiple, es igual a 0,548 lo que indica que todavía hay aproximadamente un 55 % de inseguridad en los pronósticos; elevado al cuadrado y multiplicado por 100 representa el porcentaje de varianza del criterio que no se puede explicar a partir del conjunto de variables predictoras, en nuestro caso un 30%.

El coeficiente de valor predictivo múltiple, es el complementario del coeficiente de alienación, y multiplicado por 100 representa el porcentaje de seguridad en los pronósticos, en nuestro caso un 45%.

2.4. Métodos para seleccionar las variables predictoras más adecuadas

Al hacer el análisis del puesto de trabajo es posible que se disponga de diferentes predictores para pronosticar un criterio; no obstante, antes de utilizarlos todos conviene estar seguros de que, en realidad, contribuyen de manera significativa a la predicción del criterio explicando una parte de la varianza que no es explicada por ninguno de los demás.

Para poder hacer esta selección hay varios métodos estadísticos: *Forward* (hacia adelante), *Backward* (hacia atrás). Vamos a ir explicando de forma esquemática la forma de proceder cuando se utiliza uno u otro.

2.4.1. Métodos Forward

Dentro de estos métodos vamos a explicar el más utilizado que es el *stepwise* (paso a paso):

— Se calculan las intercorrelaciones entre las distintas variables.

— Se selecciona en primer lugar la variable predictora (independiente) cuya correlación con el criterio sea más alta y se construye la ecuación de regresión.

— Se van añadiendo en la ecuación de regresión, una a una, las demás variables predictoras pero siguiendo la siguiente pauta: la segunda variable a incluir será aquella cuya correlación semiparcial con el criterio sea más alta; es decir, sea más alta después de haber eliminado de antemano el efecto que pueda estar ejerciendo dicha variable sobre la variable que se había seleccionado en primer lugar. A continuación, la tercera variable a incluir sería la que tuviera con el criterio una correlación más alta después de haber eliminado la influencia debida a la asociación entre esa variable y las otras dos seleccionadas, y así sucesivamente.

— Cada vez que se incluye una variable predictora en la ecuación de regresión se calcula el aumento que se produce en el porcentaje de varianza del criterio que explican el conjunto de variables seleccionadas (aumento en el coeficiente de determinación múltiple) y se analiza si ese aumento es estadísticamente significativo o no. El proceso se detiene cuando el aumento no es significativo.

Los paquetes estadísticos que se utilizan habitualmente, *SPSS* por ejemplo, ofrecen estos métodos.

2.4.2. Métodos Backward

Es un método inverso al anterior y menos utilizado. Al utilizar este método se procede de la siguiente manera:

— Se calcula la correlación múltiple al cuadrado (coeficiente de determinación) entre la variable criterio y todo el conjunto de predictores de que se dispone.

— Se van eliminando una a una las variables menos relevantes calculando en cada proceso de eliminación la reducción que se produce en el coeficiente de determinación.

— El proceso se detiene cuando la reducción observada sea significativa.

2.4.3. Ejemplo

Supongamos que para la predicción del éxito como piloto (Y) se cuenta con tres posibles variables predictoras: Destreza manual (X_1), Razonamiento espacial (X_2) y Control emocional (X_3), y se encarga a un psicólogo el estudio de validación correspondiente a fin de encontrar la ecuación de regresión que contribuya mejor a la predicción del criterio. La muestra de validación utilizada estuvo formada por 300 pilotos.

En la tabla adjunta (7.3) se recogen las intercorrelaciones entre las 4 variables:

TABLA 7.3
Matriz de intercorrelaciones

	Y	X_1	X_2	X_3
Y		0,80	0,75	0,86
X_1		1	0,60	0,70
X_2			1	0,65
X_3				1

— *Método Forward: stepwise*

A partir de la matriz de intercorrelaciones se selecciona en primer lugar la variable predictora cuya correlación con el criterio que se quiere predecir es más alta:

$$r_{YX_3} = 0,86$$

a continuación se calculan las correlaciones semiparciales eliminando de las variables predictoras X_1 y X_2 la influencia que pueda estar ejerciendo su relación con la variable X_3.

$$r_{Y(X_1 X_3)} = \frac{r_{YX_1} - r_{YX_3} \cdot r_{X_1X_3}}{\sqrt{1 - r^2_{X_1X_3}}} = \frac{0,80 - 0,86 \cdot 0,70}{\sqrt{1 - 0,70^2}} = \frac{0,198}{0,714} = 0,28$$

$$r_{Y(X_2 X_3)} = \frac{r_{YX_2} - r_{YX_3} r_{X_2X_3}}{\sqrt{1 - r^2_{X_2X_3}}} = \frac{0,75 - 0,86 \cdot 0,65}{\sqrt{1 - 0,65^2}} = \frac{0,191}{0,759} = 0,25$$

Dado que la correlación semiparcial más alta es $r_{Y(X_1X_3)}$, será la variable X_1 la que entre a formar parte de la ecuación de regresión en segundo lugar.

Hay que ver, sin embargo, si el aumento que experimenta la correlación múltiple al cuadrado al introducir esta segunda variable es significativo:

$$R^2_{Y \cdot X_3X_1} = r^2_{YX_3} + r^2_{Y(X_1X_3)} = 0,86^2 + 0,28^2 = 0,82$$

$$R_{Y \cdot X_3X_1} = \sqrt{0,82} = 0,90$$

Como se puede observar al introducir la nueva variable, la correlación ha pasado de 0,86 a 0,90.Para ver si el aumento ha sido significativo se utiliza el siguiente estadístico de contraste:

$$F = \left(\frac{N-K-1}{K-J}\right)\left(\frac{R^2_{Y \cdot kX} - R^2_{Y \cdot jX}}{1 - R^2_{Y \cdot kX}}\right) = \left(\frac{300-2-1}{2-1}\right)\left(\frac{0,82-0,74}{1-0,82}\right) = 132$$

donde:

N = número de sujetos de la muestra

K = número de predictores finales incluidos

j = número de predictores incluidos hasta el paso anterior

$R^2_{Y \cdot kX}$ = correlación múltiple al cuadrado con K predictores

$R^2_{Y \cdot jX}$ = correlación múltiple al cuadrado con J predictores

El estadístico de contraste tiene una distribución F de Snedecor con $(K-J)$ y $(N-K-1)$ grados de libertad.

En nuestro ejemplo, si se acude a las tablas de F (Tabla 5, al final del libro) y se busca a un determinado nivel de confianza, por ejemplo del 95%, el valor de F correspondiente a 1 y 297 grados de libertad se observa que el valor encontrado es significativo puesto que el valor obtenido en la tabla es más pequeño. Se debería introducir la variable X_1 en la ecuación de regresión.

Ya sólo queda probar si se debe introducir la variable X_2, para ello continuamos el proceso calculando las correlaciones semiparciales siguientes:

$$r_{Y(X_2 \cdot X_3X_1)} = \frac{r_{Y(X_2 \cdot X_3)} - r_{Y(X_1 \cdot X_3)} \cdot r_{X_2(X_1 \cdot X_3)}}{\sqrt{1 - r^2_{X_2(X_1 \cdot X_3)}}} = \frac{0,25 - 0,28 \cdot 0,203}{\sqrt{1 - 0,203^2}} = \frac{0,193}{0,979} = 0,197$$

$$r_{X_2(X_1 \cdot X_3)} = \frac{r_{X_1X_2} - r_{X_1X_3} \cdot r_{X_2X_3}}{\sqrt{1 - r^2_{X_1X_3}}} = \frac{0,60 - 0,70 \cdot 0,65}{\sqrt{1 - 0,70^2}} = \frac{0,145}{0,714} = 0,203$$

La correlación múltiple al cuadrado aumenta de 0,82 a 0,86 al introducir la variable X_2

$$R^2_{Y \cdot X_3X_1X_2} = r^2_{YX_3} + r^2_{Y(X_1 \cdot X_3)} + r^2_{Y(X_2 \cdot X_3X_1)} = 0,82 + 0,197^2 = 0,86$$

comprobamos si este incremento es significativo:

$$F = \frac{300-3-1}{3-2}\left(\frac{0,86-0,82}{1-0,86}\right) = 84,57$$

Acudiendo a las tablas de F con 1 y 296 grados de libertad se comprueba, al mismo nivel de confianza, que el aumento es significativo, puesto que el valor de F encontrado es mayor que el de las tablas, por lo tanto se debe incluir ésta última variable en la ecuación de regresión al contribuir a mejorar el pronóstico del criterio significativamente.

— *Método Backward*

Se procede en sentido inverso. En primer lugar se obtiene la correlación múltiple al cuadrado entre el criterio y el conjunto de variables predictoras que en nuestro ejemplo, tal y como hemos visto anteriormente, es:

$$R^2_{Y \cdot X_3 X_1 X_2} = 0,86$$

Se van eliminando una a una las variables predictoras calculando en cada caso la reducción en el coeficiente de correlación múltiple.

a) Eliminando la variable X_2, la correlación quedaría así:

$$R^2_{Y \cdot X_3 X_1} = \frac{r^2_{YX_3} + r^2_{YX_1} - 2r_{YX_3} r_{YX_1} r_{X_1 X_3}}{1 - r^2_{X_1 X_3}} =$$

$$= \frac{0,86^2 + 0,80^2 - 2 \cdot 0,86 \cdot 0,80 \cdot 0,70}{1 - 0,70^2} = 0,82$$

la reducción sería de: 0,86 – 0,82 = 0,04

b) Si se eliminara la variable X_1, la correlación sería:

$$R^2_{Y \cdot X_3 X_2} = \frac{0,75^2 + 0,86^2 - 2 \cdot 0,86 \cdot 0,75 \cdot 0,65}{1 - 0,65^2} = 0,80$$

la reducción sería : 0,86 – 0,80 = 0,06

c) Si se eliminara la variable X_3 la correlación quedaría:

$$R_{Y \cdot X_1 X_2} = \frac{0,75^2 + 0,80^2 - 2.\ 0,80 \cdot 0,75 \cdot 0,60}{1 - 0,60^2} = 0,75$$

la reducción sería: 0,86 – 0,75 = 0,11

Para ver si este decremento es significativo o no, se calcula el estadístico de contraste F como se ha ido aplicando anteriormente, los resultados son los siguientes:

$$F = 296\ \frac{0,04}{1-0,86} = 296\ \frac{0,04}{0,14} = 84,57$$

$$F = 296\ \frac{0,06}{0,14} = 126,86$$

$$F = 296\ \frac{0,11}{0,14} = 232,57$$

Se acude a las tablas de F y se busca el valor crítico para 1 y 296 grados de libertad y un nivel de confianza del 95%. Los resultados muestran que la eliminación de cualquier predictor produciría una reducción significativa en el valor de la correlación múltiple; no obstante, es la variable X_3 la que produciría una reducción más alta puesto que es la que tiene una correlación más alta con el criterio.

Nota: Aunque esta introducción a los métodos de selección de los predictores se ha hecho de manera muy esquemática, creemos que puede ayudar a nuestros alumnos a comprender el proceso.

3. VALIDEZ Y UTILIDAD DE LAS DECISIONES

Se incluyen en este apartado una serie de procedimientos que van a permitir analizar la validez de las decisiones tomadas a partir de las puntuaciones obtenidas por los sujetos en un test (o varios) en relación a un criterio dicotómico. Pero, a diferencia de lo que ocurriría si las variables predictoras fueran variables cuantitativas y el criterio dicotómico, donde el procedimiento estadístico más adecuado para analizar la validez de las inferencias sería la regresión logística, la situación que se plantea ahora es que las puntuaciones obtenidas en el test se dicotomizan a partir de un punto de corte de manera que permitan asignar a los sujetos en dos categorías, por ejemplo, *Admitidos-Rechazados* en un puesto de trabajo, *Aptos-No aptos* en un examen, *Enfermos-No enfermos*, etc.

En este tipo de situaciones, no tendría demasiado sentido estudiar la capacidad predictiva del test mediante coeficientes de correlación como los utilizados anteriormente, sino mediante unos índices que reflejen la consistencia o acuerdo entre las decisiones basadas en el test y la medida del criterio. Estos procedimientos son los que se utilizan generalmente en los Tests Referidos al Criterio (TRC), tal y como se ha expuesto en el tema 5, y en muchas situaciones aplicadas.

3.1. Índices de validez y de selección

Para una mejor comprensión de la lógica del proceso vamos a utilizar un ejemplo.

EJEMPLO:

Supongamos que se quiere llevar a cabo la selección de los alumnos que van a hacer el Doctorado en el Departamento de Metodología de las Ciencias del Comportamiento de la UNED el próximo curso y no sabemos si la prueba de admisión con la que contamos puede servir a nuestros propósitos. Para ello, vamos a llevar a cabo un estudio de validación. Se aplica la prueba a todos los que han presentado su solicitud para este año, y dado que se exigen unos conocimientos mínimos para poder tener acceso a los cursos y que se desea seleccionar a los mejores, se fija un punto de corte (X_c), de manera que todos aquellos sujetos que obtengan puntuaciones por encima del punto de corte serán considerados aptos (*A*) para hacer el doctorado y los que no lo alcancen serán considerados no aptos (*R*). En este caso, la prueba utilizada como predictor para tomar decisiones acerca de la adecuación o no de los aspirantes a realizar el doctorado en nuestro Departamento es una variable dicotomizada (puntuaciones por encima o por debajo del punto de corte). Se admite a todos los aspirantes en los cursos de Doctorado, y al finalizar el año se les evalúa de manera que la calificación asignada permita diferenciar entre aquellos alumnos que han tenido un buen rendimiento en los cursos de Doctorado (*A*) y los que no han tenido un buen rendimiento (*R*). Diremos que la prueba utilizada será válida para hacer la selección de los alumnos de doctorado, y por lo tanto se podrá utilizar en la selección del curso siguiente, cuando las decisiones tomadas a partir de las puntuaciones en el test se vean confirmadas con las decisiones tomadas en base a los criterios de rendimiento marcados por el Departamento.

Supongamos que los resultados del proceso de validación son los que aparecen recogidos en la tabla 7.4.

TABLA 7.4
Clasificación de los sujetos en función del test y del criterio

		Criterio A	Criterio R	
Test	A	N_{AA} (18)	N_{AR} (2)	N_{AT} (20)
	R	N_{RA} (3)	N_{RR} (27)	N_{RT} (30)
		N_{AC} (21)	N_{RC} (29)	N (50)

En la tabla anterior :

$N_{AA} + N_{RR}$ = (Aciertos).

Número de alumnos que han sido calificados del mismo modo en la prueba de selección (test) y en el criterio. Los primeros han sido considerados aptos tanto en la prueba de admisión como en el criterio y los segundos han sido rechazados en ambas calificaciones.

N_{RA} = (Falsos negativos).

Alumnos que superaron el criterio de rendimiento y sin embargo en la prueba de admisión no superaron el punto de corte. En un proceso de selección habrían sido rechazados y, sin embargo, deberían haber sido admitidos.

N_{AR} = (Falsos positivos).

Alumnos que en la prueba de admisión superaron el punto de corte y luego no superaron el criterio de rendimiento. En un proceso de selección no deberían haber sido seleccionados y, sin embargo, al superar el punto de corte en el predictor serían admitidos.

N_{AC} = número de alumnos que han sido considerados aptos en el criterio.

N_{RC} = número de alumnos que han sido considerados no aptos en el criterio.

N_{AT} = número de alumnos que han sido considerados aptos en el test.

N_{RT} = número de alumnos que han sido considerados no aptos en el test.

3.1.1. Índices de validez

— *Coeficiente Kappa*

A partir de los datos es necesario obtener algún indicador de la validez de la prueba de admisión para pronosticar el criterio, uno de los más utilizados es el *Coeficiente Kappa* de Cohen (1960)

que permite evaluar la consistencia o acuerdo entre las decisiones adoptadas a partir de las puntuaciones obtenidas por los sujetos en el predictor (en nuestro ejemplo la prueba de admisión) y en el criterio (en nuestro ejemplo el rendimiento en el doctorado).

La fórmula del coeficiente viene dada por:

$$K = \frac{F_c - F_a}{N - F_a} \qquad [7.16]$$

donde:

F_c = número de casos en los que hay coincidencia entre las puntuaciones del predictor y las del criterio.

F_a = número de casos en los que cabe esperar que las calificaciones del predictor y las del criterio coincidan por azar.

N = número de personas de la muestra.

Para calcular las frecuencias esperadas por azar, F_a, se multiplican las frecuencias marginales correspondientes y se dividen por el número total de sujetos.

En nuestro ejemplo:

$$F_c = N_{AA} + N_{RR} = (18 + 27) = 45$$

Para calcular F_a se procede de la siguiente manera:

$$\text{Frecuencia esperada de la casilla } AA = \frac{21 \times 20}{50} = 8,4$$

$$\text{Frecuencia esperada de la casilla } RR = \frac{29 \times 30}{50} = 17,4$$

$$F_a = 8,4 + 17,4 = 25,8$$

$$K = \frac{45 - 25,8}{50 - 25,8} = \frac{19,2}{24,2} = 0,79$$

Puesto que el valor máximo del coeficiente *Kappa* es 1, la validez de la prueba de admisión para pronosticar el criterio de rendimiento es alta. Ante estos resultados se podría utilizar la prueba en cursos posteriores para hacer la selección de los alumnos que quieren hacer el Doctorado en el Departamento.

Del análisis de la tabla 7.4 se puede obtener más información para valorar los resultados de la decisión adoptada.

— *Proporción de clasificaciones correctas*

$$P.C.C. = \frac{N_{AA} + N_{RR}}{N} = \frac{18 + 27}{50} = 0,90$$

— *Sensibilidad*

Es un índice que equivale a la proporción de aspirantes correctamente seleccionados mediante la prueba de admisión respecto al total de los que tuvieron éxito en el criterio, es decir, respecto al total de los sujetos que rindieron satisfactoriamente en los cursos de doctorado del Departamento. De los 21 aspirantes que tuvieron un rendimiento adecuado en los cursos de doctorado, 18 habían sido detectados mediante la prueba de admisión.

$$S = \frac{N_{AA}}{N_{AC}} = \frac{18}{21} = 0,86$$

— *Especificidad*

Proporción de aspirantes que fueron correctamente rechazados mediante la prueba de admisión respecto al total de los aspirantes que no alcanzaron un rendimiento adecuado en los cursos del doctorado. De los 29 aspirantes que no tuvieron un rendimiento satisfactorio en los cursos de doctorado, 27 habían sido detectados mediante la prueba de admisión.

$$E = \frac{N_{RR}}{N_{RC}} = \frac{27}{29} = 0,93$$

Dado que el valor máximo de estos índices es la unidad, se puede decir que la prueba de admisión tiene una buena capacidad predictiva.

— *Razón de Eficacia*

Proporción de aspirantes seleccionados mediante la prueba de admisión que rindieron satisfactoriamente en el doctorado.

$$R.E. = \frac{N_{AA}}{N_{AT}} = \frac{18}{20} = 0,90$$

3.1.2. Índices de selección

Además de los índices de validez, en un proceso de selección se puede obtener otros índices que ofrecen información acerca del resultado del proceso:

— *Razón de Idoneidad*

Cuando se lleva a cabo una selección, la razón de idoneidad equivale a la proporción de aspirantes que rindieron satisfactoriamente en el criterio.

$$R.I. = \frac{N_{AC}}{N} = \frac{21}{50} = 0,42$$

— *Razón de Selección*

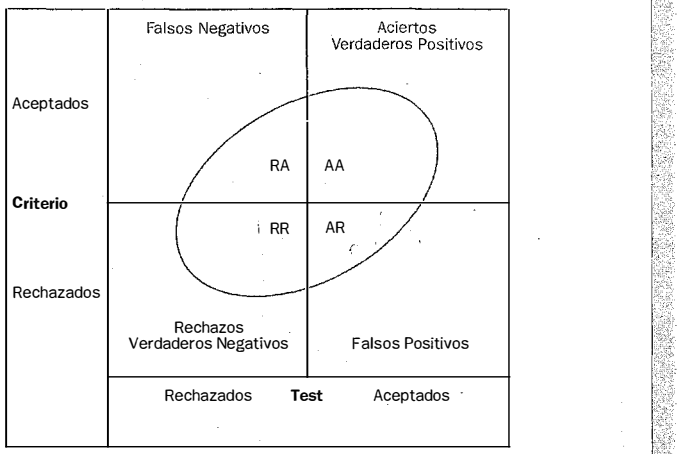
En un proceso de selección, es la proporción de aspirantes que han sido seleccionados mediante el test.

$$R.S. = \frac{N_{AT}}{N} = \frac{20}{50} = 0,40$$

3.2. ¿Donde situar el punto de corte?

Si, como hemos comentado en el punto anterior, para seleccionar a los sujetos mediante una prueba de admisión (variable predictora), y clasificarlos en las dos categorías de: Admitidos- Rechazados, era necesario dicotomizar las puntuaciones obtenidas estableciendo un punto de corte, es fácil darse cuenta de la importancia que tiene el valor correspondiente a ese punto de corte sobre la validez de la prueba. Por otra parte, también es importante el punto de corte del criterio ya que es el que nos va a permitir definir las dos categorías de rendimiento: Satisfactorio-No satisfactorio. La figura 7.1 ayudará a comprender todo lo dicho hasta el momento:

FIGURA 7.1
Resultados de las decisiones



Vamos a suponer que la elipse representa el diagrama de la distribución conjunta de las puntuaciones obtenidas por los sujetos de la muestra (en nuestro ejemplo, por los aspirantes al curso de doctorado) tanto en el test predictor cuya validez se quiere probar (prueba de admisión) como en el criterio (rendimiento en el curso de doctorado). Las dos líneas que se cruzan, y que dividen a la elipse en 4 partes, corresponden a los puntos de corte establecidos tanto en el test como en el criterio.

A partir de la figura se puede comprender la importancia de situar los puntos de corte en el lugar adecuado. Si el punto de corte del test se moviera hacia la derecha, es decir, si se utilizara un criterio de selección más estricto, se reduciría la tasa de falsos positivos (AR) y por lo tanto habría un menor número de aspirantes que habiendo sido seleccionados por el test no alcanzaran el rendimiento adecuado en el criterio; ahora bien, habría también un aumento en la tasa de falsos negativos (RA), lo que implicaría que se quedarían fuera un mayor número de aspirantes que, si hubieran sido seleccionados, podrían haber rendido adecuadamente en el criterio.

Si es el criterio el que se hace más estricto, disminuirá el número de falsos negativos (RA) pero a costa de que aumente el número de falsos positivos (AR).

Entonces, ¿dónde se debe situar el punto de corte? Dado que la validez de las decisiones que se tomen va a depender de donde se sitúe el punto de corte, habrá que buscar el valor de éste que

maximice la capacidad predictiva de la variable predictora. En este sentido el punto de corte debería situarse, en principio, en el punto que hiciera mínimos los errores de clasificación. Pero, por otra parte, hay que analizar las consecuencias de las decisiones tomadas ya que no siempre tiene la misma importancia cometer un tipo de error u otro (falsos positivos o falsos negativos), ello dependerá del tipo de decisión a tomar, por eso éste es otro factor a tener en cuenta a la hora de situar el punto de corte.

Si se hablara en términos de la teoría de la decisión estadística diríamos que el punto de corte habría que situarlo teniendo en cuenta la matriz de pagos, es decir, la matriz que refleje las «*pérdidas*» y «*ganancias*» derivadas de las decisiones adoptadas. Sin entrar en este tema ya que queda fuera de nuestros objetivos, señalar dos criterios que se suelen utilizar cuando se han de tomar decisiones *en ambiente de incertidumbre*, es decir, sin saber cuál va a ser el resultado exacto de la decisión, nos referimos al *Criterio maximin* y al *Criterio minimax*.

Cuando un decisor utiliza un criterio maximin, deberá elegir aquella alternativa que entre los resultados más desfavorables, le permita obtener la máxima ganancia (máximo de los mínimos). Cuando utiliza el criterio mínimax, el decisor estudiaría las alternativas que le van a proporcionar las máximas «pérdidas» y dentro de esas alternativas elegiría aquella que le proporcionara una «pérdida» menor (mínimo de los máximos).

3.3. Ejemplo

Siguiendo con el ejemplo de la selección de alumnos al curso de doctorado del Departamento, vamos a suponer que las calificaciones obtenidas en la prueba de admisión y en el criterio de rendimiento por un grupo de 10 aspirantes fueron las que figuran en la tabla 7.5. Si se considerara que para poder ser admitido a los cursos de Doctorado los aspirantes deberían haber obtenido una calificación de 7 puntos o más en la prueba de admisión ¿Cuál sería la validez de la prueba para predecir el rendimiento en los cursos de Doctorado?

TABLA 7.5
Datos del ejemplo

Aspirantes	Prueba	Criterio
A	5	NA
B	7	NA
C	6	A
D	8	A
E	6	NA
F	7	A
G	6	A
H	9	A
I	4	NA
J	6	NA

Si se considera que la prueba de admisión es una variable cuantitativa y el criterio es una variable dicotómica (NA = No apto, y A = Apto), para estimar la validez de la prueba en relación al criterio se podría calcular la correlación biserial puntual entre las puntuaciones obtenidas en ambas variables y obtener así el coeficiente de validez. Si se estableciera la dicotomía de la variable predictora mediante el punto de corte ($X \geq 7$), se tendría una variable dicotomizada y una variable dicotómica, en este caso el coeficiente más adecuado sería la correlación «fi-biserial». Cualquiera de estas opciones sería correcta, pero la información que ofrecieran sería muy general ya que no se podría saber nada acerca de los errores cometidos al hacer la selección que, a nuestro juicio, son importantes. Es preferible proceder de la siguiente manera:

Se elabora una tabla de contingencia de 2 × 2 (tabla 7.6) en la que se reflejen las decisiones conjuntas tomadas a partir de la prueba de admisión y del rendimiento en el criterio:

TABLA 7.6
Distribución de aspirantes

		Prueba de admisión		
		Aceptado ≥ 7	Rechazado < 7	
Criterio	Apto	3 (Aciertos)	2 (Falso negativo)	5
	No Apto	1 (Falso positivo)	4 (Aciertos)	5
		4	6	10

De la distribución de los aspirantes en la tabla se puede sacar la siguiente información:

- Hay 3 aspirantes (D, F y H) que han superado el punto de corte en la prueba de admisión y, además, han tenido un buen rendimiento en el Doctorado (ACIERTOS).
- Hay 4 aspirantes (A, E, I y J) que han sido también correctamente clasificados ya que no superaron la prueba de admisión y, a su vez, tuvieron un mal rendimiento en el Doctorado (ACIERTOS).
- Hay una persona que alcanzó en la prueba de admisión la puntuación necesaria para ser admitido y, sin embargo, luego tuvo un mal rendimiento en el Doctorado (B) (FALSO POSITIVO).
- Hay 2 personas (C y G) que no habiendo alcanzado la puntuación mínima necesaria en la prueba de admisión, sí que rindieron bien en el Doctorado (FALSOS NEGATIVOS).

Con estos datos se pueden obtener los índices de validez y de selección que se han explicado anteriormente:

— *Índice Kappa*

$$K = \frac{F_c - F_a}{N - F_a} = \frac{7-5}{10-5} = 0,40$$

$$F_a(AA) = \frac{4 \times 5}{10} = 2 \qquad F_a(RR) = \frac{6 \times 5}{10} = 3 \qquad F_a = 2 + 3 = 5$$

— *Proporción de clasificaciones correctas*

$$PCC = \frac{AA + RR}{N} = \frac{3+4}{10} = 0,70$$

— *Sensibilidad*

$$S = \frac{AA}{AC} = \frac{3}{5} = 0,60$$

— *Especificidad*

$$E = \frac{RR}{RC} = \frac{4}{5} = 0,80$$

— *Razón de idoneidad*

$$RI = \frac{AC}{N} = \frac{5}{10} = 0,50$$

— *Razón de eficacia*

$$RE = \frac{AA}{AT} = \frac{3}{4} = 0,75$$

Teniendo en cuenta que el valor máximo que se puede obtener en cada uno de los índices es la unidad, los valores obtenidos son bastante aceptables.

3.4. Modelos de Selección

Ya se comentó anteriormente que a la hora de tomar decisiones acerca de la competencia o no de una muestra de sujetos para desarrollar un trabajo, del rendimiento de los alumnos en determinados programas, de la adscripción de un grupo de pacientes a un determinado tipo de terapia, etc., es necesario obtener el máximo de información para evitar cometer errores que, de otra manera, se hubieran podido evitar. En general, esta información se obtiene a partir del *currículum vitae,* de las puntuaciones obtenidas en ciertos tests, mediante entrevistas, dinámicas de grupo, etc., pero el problema que surge es cómo combinar toda esa información a la hora de tomar una decisión.

Hay tres modelos básicos a los que se pueden añadir dos de tipo mixto (los dos últimos):

- Compensatorio
- Conjuntivo
- Disyuntivo
- Conjuntivo-compensatorio
- Disyuntivo-compensatorio

— *Modelo compensatorio*

Se trata de un modelo aditivo en el que a cada sujeto se le asigna una única puntuación global. El nombre alude a que los sujetos pueden compensar una baja puntuación en una de las pruebas con una puntuación alta en otras de manera que el resultado final sea una única puntuación (por ejemplo el examen de selectividad). Este tipo de modelo no siempre tiene sentido ya que hay veces que la ausencia de alguna destreza o capacidad no puede ser compensada por un exceso en

otra. Si un requisito imprescindible para un puesto de trabajo es el conocimiento de la lengua inglesa, difícilmente se podrá compensar una falta de conocimiento de este idioma con una buena puntuación en una prueba de conocimientos informáticos.

Una forma adecuada de obtener la puntuación global, a partir de la combinación aditiva de todas las puntuaciones obtenidas en las distintas pruebas utilizadas, es mediante el modelo de regresión lineal múltiple, que ya expusimos anteriormente. Este modelo permite asignar a cada sujeto una única puntuación (la puntuación pronosticada), a partir de una combinación aditiva de los resultados obtenidos en los diferentes predictores, asignando a cada predictor un determinado peso que vendrá determinado por el coeficiente de regresión correspondiente.

— *Modelo conjuntivo*

En este modelo se fijan de antemano unos mínimos en cada una de las pruebas utilizadas para la selección, de manera que sólo se seleccionarán aquellas personas que hayan superado esos mínimos en todas y cada una de las pruebas.

— *Modelo disyuntivo*

En este modelo sólo se exige superar determinado nivel de competencia en al menos alguno de los predictores o bloque de predictores.

— *Modelo conjuntivo-compensatorio*

Se aplica, en un primer momento, el modelo conjuntivo y se seleccionan aquellos sujetos que superan los mínimos establecidos en cada uno de los predictores. A continuación, a los sujetos seleccionados se les aplica el modelo compensatorio de manera que queden ordenados en función de la puntuación global obtenida. Una vez ordenados los sujetos, dependiendo de cómo se haya planteado el proceso de selección, se puede elegir a un determinado número de entre los mejores, o bien establecer un punto de corte de manera que sean seleccionados aquellos cuya puntuación global supere el punto establecido.

— *Modelo disyuntivo-compensatorio*

Se hace una primera selección aplicando el modelo disyuntivo y a los sujetos seleccionados se les aplica el modelo compensatorio.

3.5. ¿Cómo estimar la eficacia de una selección?

Entre los índices que hemos expuesto anteriormente uno de ellos es la razón de eficacia que representa la proporción de personas seleccionadas que tienen éxito en el criterio.

Otra forma de estimar la eficacia de la selección es utilizando el modelo de regresión, siempre que se verifiquen los supuestos que implica, pues permite estimar la probabilidad de que los seleccionados tengan éxito en el criterio.

Se pueden presentar varias situaciones, pero vamos a estudiar sólo dos. Una, aquella en la que no hay un número limitado de plazas y se seleccionan todas aquellas personas que superan una determinada puntuación en el predictor (o predictores) y la otra situación es aquella en la que sí hay un número de plazas limitadas y se quiere seleccionar a los que hayan obtenido mejores resultados en el predictor (o predictores).

EJEMPLO:

Supongamos que la ecuación de regresión obtenida a partir de un test (X) para predecir un criterio (Y) ha sido: $Y' = 0{,}5 + 2X$, que la desviación típica del criterio es $S_y = 5$, que el coeficiente de validez es $r_{xy} = 0{,}80$ y que para considerar que se ha tenido éxito en el criterio es necesario obtener en el mismo una puntuación igual o mayor de 8 puntos. Con estos datos, y suponiendo que no hay un número limitado de plazas, ¿qué probabilidad de éxito tendrán los sujetos que en el test hayan obtenido una puntuación de 6 puntos?

— En primer lugar se estima la puntuación pronosticada en el criterio de los sujetos que en el test obtuvieron una puntuación de 6 puntos. Esta puntuación es la media de la distribución de todas las puntuaciones que han podido obtener en el criterio los sujetos que en el test obtuvieron 6 puntos. La desviación típica de esa distribución es el error típico de estimación:

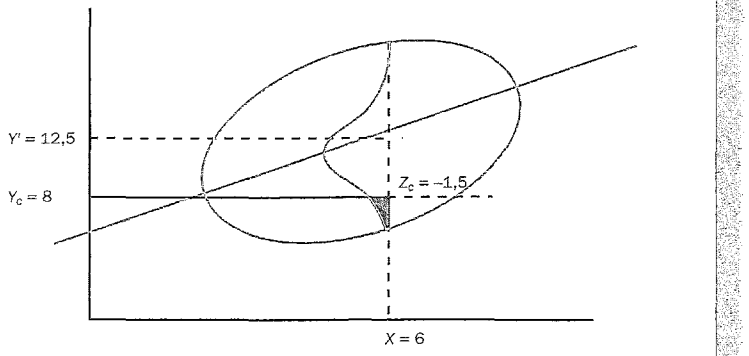
$$Y' = 0{,}5 + 2(6) = 12{,}5$$

— Se asume que las distribuciones de los errores de estimación condicionados a una determinada puntuación en el test se ajustan a la distribución normal de probabilidad, con una media que vendrá dada por la puntuación pronosticada en el criterio y con la desviación típica que vendrá dada por el error típico de estimación. Así pues es necesario calcular el error típico de estimación:

$$S_{Y \cdot X} = S_Y \sqrt{1 - r_{XY}^2} = 5\sqrt{1 - 0{,}64} = 3$$

La puntuación típica correspondiente al punto crítico del criterio Z_c es la que va a marcar la separación entre la probabilidad de éxito y la de fracaso y, por lo tanto, la que nos va a permitir analizar la eficacia de la selección. Su cálculo se hace así:

FIGURA 7.2



$$Z_C = \frac{Y_C - Y'}{S_{Y \cdot X}} = \frac{8 - 12,5}{3} = -1,5$$

Se acude a las tablas de curva normal (al final del libro) y se busca el área de la curva que queda por encima de una puntuación típica de –1,5. Obtenemos una probabilidad de 0,9332; luego un sujeto que obtuviera en el test 6 puntos y haya sido seleccionado al pronosticársele en el criterio una puntuación de 12,5 que está por encima del punto crítico, tiene una probabilidad de 0,9332 de tener éxito. La probabilidad de fracaso sería 1 – 0,9332 = 0,0668 (zona oscura de la figura).

Siguiendo con los datos del ejemplo anterior, vamos a hacer un planteamiento distinto. Supongamos que hay 100 aspirantes al puesto de trabajo y que sólo hay 10 plazas a cubrir. En este caso se supone que deberán seleccionarse a los 10 aspirantes que hayan obtenido mejores puntuaciones en el test. ¿Cuál sería la probabilidad de éxito de estas personas?

En primer lugar hemos de averiguar cuál es la puntuación mínima que han obtenido las 10 personas seleccionadas por el test. Como sabemos que esas personas representan el 10% de todos los aspirantes, deberemos buscar la puntuación que deja por debajo el 90% de la muestra de aspirantes. Si asumimos que las puntuaciones en el test se distribuyen según la curva normal de probabilidad, la puntuación típica que deja por debajo el 90% de los casos es $Z_X = 1,28$ aproximadamente. Si la media del test fuera de 7 puntos y la desviación típica de 2 puntos, la puntuación directa mínima de las 10 personas seleccionadas sería:

$$1,28 = \frac{X - \bar{X}}{S_X} = \frac{X - 7}{2} \Rightarrow X = 7 + 1,28(2) = 9,56$$

Una vez obtenida esta puntuación procedemos de la misma manera que en el ejemplo anterior, se aplica la ecuación de regresión y se calcula la puntuación que se les pronosticaría a estos sujetos en el criterio.

$$Y' = 0,5 + 2\,(9,56) = 19,62$$

Conocida la puntuación pronosticada se calcula la Z_C a partir de la cual se averigua la probabilidad de éxito de estos sujetos.

$$Z_C = \frac{8 - 19,62}{3} = -3,87$$

No es necesario acudir a las tablas de curva normal para darnos cuenta de que la probabilidad de que todos los seleccionados tengan éxito es prácticamente el 100%.

4. FACTORES DE INFLUYEN EN EL COEFICIENTE DE VALIDEZ

Son varios los factores que influyen en el valor del coeficiente de validez, pero vamos a señalar tres que a nuestro juicio son decisivos: a) la variabilidad de la muestra, b) la fiabilidad de las puntuaciones del test y del criterio y c) la longitud del test.

4.1. La variabilidad de la muestra

El coeficiente de validez se ha definido como la correlación entre las puntuaciones obtenidas por los sujetos en el predictor (o predictores) y las obtenidas en el criterio, y como tal correlación tiende a aumentar a medida que la variabilidad de la muestra utilizada es mayor y, por el contrario, tiende a disminuir a medida que la muestra es más homogénea. Por lo tanto, el conocimiento de la variabilidad de la muestra es fundamental a la hora de poder interpretar el coeficiente de validez, ya que para un mismo predictor y una misma medida del criterio el coeficiente puede variar de muestra a muestra.

Dado que lo que nos interesa es que nuestros alumnos comprendan de qué manera influye la variabilidad de la muestra en el coeficiente de validez vamos a exponer con un ejemplo el caso

más sencillo, aquél en el que sólo hay dos variables implicadas, una la variable predictora y otra la variable criterio.

EJEMPLO:

Supongamos que una Universidad privada utiliza, además de otras técnicas, una batería de tests para hacer la selección de sus alumnos. Si se quiere conocer la validez de esa batería para pronosticar el rendimiento de los alumnos en sus estudios, será necesario buscar algún indicador que permita obtener una medida de ese rendimiento; un indicador puede ser las notas obtenidas al finalizar el primer año de licenciatura. Para averiguar la validez predictora de la batería utilizada se calculará la correlación entre las puntuaciones obtenidas por los sujetos en la batería y la medida del criterio.

El valor obtenido será el coeficiente de validez de la batería, pero se ha obtenido en una muestra previamente seleccionada puesto que las calificaciones en el criterio sólo se conocen en la muestra de admitidos. Esta muestra será mucho más homogénea que la formada por todos los aspirantes y, por lo tanto, el valor de la correlación obtenida será más bajo.

Dado que lo que en realidad interesa es conocer la capacidad predictora de la batería antes de hacer la selección, es decir en el grupo de aspirantes, ya que no tendría sentido seleccionar a un grupo de sujetos y que luego se pusiera de manifiesto que la batería no servía para predecir el criterio elegido, hay dos formas de proceder:

a) Aplicar la batería a todos los aspirantes, admitirlos a todos, y al finalizar el primer año de su carrera evaluarles en el criterio de rendimiento académico a partir de las notas que hubieran obtenido. La correlación entre las puntuaciones obtenidas en la batería de tests y las notas obtenidas sería el coeficiente de validez de la batería. Creo que si este fuera el método necesario para llevar a cabo el proceso de validación de la batería, se utilizarían otras técnicas para hacer la selección.

b) Una forma alternativa de llevar a cabo el proceso de validación es, tal y como se ha comentado anteriormente, calcular la correlación entre las puntuaciones obtenidas en la batería por el grupo de alumnos seleccionados y sus puntuaciones en el criterio y, posteriormente, basándose en una serie de supuestos hacer una estimación del coeficiente de validez que se habría obtenido en el grupo de aspirantes.

— *Supuestos*

— La pendiente de la ecuación de regresión que permitirá pronosticar el criterio a partir de la variable predictora es la misma en el grupo de aspirantes y en el de seleccionados.

— El error típico de estimación es igual en ambos grupos

Si denotamos con letras mayúsculas los datos referidos al grupo de aspirantes y con minúsculas los del grupo de admitidos, la expresión formal de estos supuestos será:

$$B = b \Rightarrow R_{XY}\frac{S_Y}{S_X} = r_{xy}\frac{s_y}{s_x}$$

$$S_{Y\cdot X} = s_{y\cdot x} \Rightarrow S_Y\sqrt{1-R_{XY}^2} = s_y\sqrt{1-r_{xy}^2} \qquad [7.17]$$

Si lo que se desea es conocer el coeficiente de validez de la batería en el grupo de aspirantes, basta despejarlo de las dos ecuaciones anteriores

$$R_{XY} = \frac{S_X \cdot r_{xy}}{\sqrt{S_X^2 \cdot r_{xy}^2 + s_x^2 - s_x^2 \cdot r_{xy}^2}} \qquad [7.18]$$

Si se quiere estimar cuál sería la variabilidad de la muestra de aspirantes en el criterio, bastaría con despejar S_y de las ecuaciones anteriores:

$$S_Y = s_y\sqrt{1-r_{xy}^2 + r_{xy}^2\frac{S_X^2}{s_x^2}} \qquad [7.19]$$

Vamos a suponer que el número de aspirantes era 300 y obtuvieron una desviación típica en la batería de 12 puntos. De entre todos ellos se seleccionaron a 40, cuya desviación típica en la batería fue de 6 puntos. Al cabo del año los admitidos fueron calificados en el criterio, siendo la correlación entre las puntuaciones que habían obtenido en la batería y las del criterio 0,30.

¿Cuál sería el coeficiente de validez estimado si se hubiese calculado en la muestra total de aspirantes?

$$R_{XY} = \frac{(12)(0,30)}{\sqrt{(12)^2(0,30)^2 + 6^2 - 6^2(0,30)^2}} = \frac{3,6}{6,76} = 0,53$$

Se puede apreciar que el cambio es bastante grande y, sin embargo, la desviación típica sólo ha pasado de 12 a 6 puntos. Si hubiera habido más diferencia entre las dos desviaciones típicas el cambio hubiera sido aún mayor.

Aunque es poco probable se puede dar el caso contrario, que se conozcan los datos en el grupo de aspirantes y se quisiera conocer cuál sería el coeficiente de validez en el grupo de seleccionados.

Teniendo en cuenta los dos supuestos de los que se parte, y de los que se derivan todas las fórmulas, en lugar de despejar R_{XY} despejaríamos r_{XY}, la fórmula resultante sería igual que la anterior pero cambiando las letras minúsculas por mayúsculas y viceversa.

4.2. La fiabilidad de las puntuaciones del test y del criterio

Cuando se calcula el coeficiente de validez como la correlación entre las puntuaciones empíricas obtenidas por los sujetos en el test y en el criterio hay que tener en cuenta que esas puntuaciones empíricas están afectadas por errores de medida y que esos errores de medida están influyendo en el coeficiente de validez produciendo una serie de sesgos que es necesario eliminar o, al menos, controlar. Spearman (1904) propuso una fórmula a la que denominó *fórmula de atenuación* porque permite corregir la *atenuación*, disminución o reducción del coeficiente de validez debida a la presencia de los errores de medida. De esta fórmula se pueden derivar varios casos particulares que van a ser analizados con un ejemplo.

EJEMPLO:

Aplicado un test de razonamiento abstracto a una muestra de sujetos se obtuvo un coeficiente de fiabilidad igual a 0,64, la fiabilidad del criterio resultó ser 0,60 y el coeficiente de validez 0,56.

4.2.1. Estimación del coeficiente de validez en el supuesto de que tanto el test como el criterio tuvieran una fiabilidad perfecta

La fórmula viene expresada por:

$$R_{V_X V_Y} = \frac{r_{XY}}{\sqrt{r_{XX} \cdot r_{YY}}} \qquad [7.20]$$

donde:

$R_{V_X V_Y}$ = coeficiente de validez teórico que se obtendría si las puntuaciones del test y del criterio estuvieran libres de errores de medida. En este caso la correlación se calcularía entre las puntuaciones verdaderas del test y del criterio.

r_{XY} = coeficiente de validez empírico.

r_{XX} = coeficiente de fiabilidad empírico del test.

r_{YY} = coeficiente de fiabilidad empírico del criterio.

¿De dónde surge esta fórmula?

La correlación entre las puntuaciones verdaderas en el test y las verdaderas en el criterio sería igual a la covarianza entre ambas dividida por el producto de las desviaciones típicas de las puntuaciones verdaderas de ambos:

$$R_{V_X V_Y} = \frac{Cov(V_X V_Y)}{S_{V_X} S_{V_Y}} = \frac{Cov(XY)}{S_{V_X} S_{V_Y}} = \frac{r_{XY} S_X S_Y}{S_{V_X} S_{V_Y}} = \frac{r_{XY}}{\frac{S_{V_X} S_{V_Y}}{S_X S_Y}} = \frac{r_{XY}}{r_{VX} r_{VY}} = \frac{r_{XY}}{\sqrt{r_{XX}}\sqrt{r_{YY}}}$$

Una de las deducciones del modelo de Spearman es que la covarianza entre las puntuaciones verdaderas es igual a la de las empíricas, por eso se sustituye en la fórmula; pero, además, teniendo en cuenta que la covarianza es igual a la correlación de Pearson entre las dos series de puntuaciones multiplicada por las dos desviaciones típicas, que el cociente entre la desviación típica de las puntuaciones verdaderas y la de las empíricas es el índice de fiabilidad y que éste es la raíz cuadrada del coeficiente de fiabilidad, se obtiene la fórmula propuesta.

¿Cuál sería el coeficiente de validez estimado en el caso de que tanto las puntuaciones del test como las del criterio estuvieran libres de errores de medida?

$$R_{V_X V_Y} = \frac{0,56}{\sqrt{0,64}\sqrt{0,60}} = \frac{0,56}{0,80 \cdot 0,77} = 0,91$$

Como se puede observar si se pudieran eliminar todos los errores de medida que afectan a las puntuaciones del test y del criterio habría un aumento considerable del coeficiente de validez, pasaría de 0,56 a 0,91.

4.2.2. Estimación del coeficiente de validez en el supuesto de que el test tuviera una fiabilidad perfecta

Partiendo de la fórmula anterior, supongamos que ahora sólo el test tiene fiabilidad perfecta. En este caso la estimación del coeficiente de validez se haría calculando la correlación entre las puntuaciones verdaderas del test y las empíricas del criterio.

$$R_{V_X Y} = \frac{r_{XY}}{\sqrt{r_{XX}}} \qquad [7.21]$$

Para la deducción de la fórmula se sigue el mismo razonamiento:

$$R_{V_X Y} = \frac{Cov(V_X Y)}{S_{V_X} S_Y} = \frac{Cov(XY)}{S_{V_X} S_Y} = \frac{r_{XY} S_X S_Y}{S_{V_X} S_Y} = \frac{r_{XY}}{\frac{S_{V_X} S_Y}{S_X S_Y}} = \frac{r_{XY}}{r_{VX}} = \frac{r_{XY}}{\sqrt{r_{XX}}}$$

Si tomamos otra vez el ejemplo, la estimación del coeficiente de validez sería:

$$R_{V_XY} = \frac{0,56}{\sqrt{0,64}} = \frac{0,56}{0,80} = 0,70$$

El valor del coeficiente de validez aumenta con respecto al valor inicial, pero este aumento, aunque grande, es más moderado que en el caso anterior ya que sólo se han eliminado los errores de medida de una de las variables (el test) pero no del criterio, cuyas puntuaciones continúan afectadas por los errores.

4.2.3. Estimación del coeficiente de validez en el supuesto de que el criterio tuviera una fiabilidad perfecta

$$R_{XV_Y} = \frac{Cov(XV_Y)}{S_X S_{V_Y}} = \frac{Cov(XY)}{S_X S_{V_Y}} = \frac{r_{XY} S_X S_Y}{S_X S_{V_Y}} = \frac{r_{XY}}{\frac{S_X S_{V_Y}}{S_X S_Y}} = \frac{r_{XY}}{r_{YY}} = \frac{r_{XY}}{\sqrt{r_{YY}}} \qquad [7.22]$$

Es el mismo procedimiento que en el caso anterior pero ahora es el criterio el que está libre de errores de medida.

Aplicando la fórmula a los datos del ejemplo tendríamos:

$$R_{XV_Y} = \frac{0,56}{\sqrt{0,60}} = 0,73$$

Vemos también que el coeficiente de validez aumentaría considerablemente a pesar de que todavía el test está afectado de errores de medida.

Estos tres casos son hipotéticos ya que, en la práctica, nunca se va a conseguir eliminar por completo los errores de medida del test, del criterio o de ambos. No obstante, sin llegar a eliminarlos del todo, sí que es posible tratar de reducirlos de alguna manera y conocer cuál sería el cambio experimentado por el coeficiente de validez en cada caso. Los tres casos que se presentan a continuación nos explican cómo hacerlo.

4.2.4. Estimación del coeficiente de validez del test en el supuesto de que se mejorara la fiabilidad tanto del test como del criterio

En la fórmula las letras mayúsculas corresponden a los coeficientes de fiabilidad mejorados.

$$R_{XY} = \frac{r_{XY}}{\sqrt{\frac{r_{XX}}{R_{XX}}}\sqrt{\frac{r_{YY}}{R_{YY}}}} \qquad [7.23]$$

Para derivar esta fórmula se parte del siguiente razonamiento: Aunque se lograse mejorar la fiabilidad del test y del criterio, eliminando en parte los errores de medida, lo que se mantendría constante sería la correlación entre las puntuaciones verdaderas del test y del criterio ya que estas puntuaciones están libres de errores. Una vez hecho esto se igualan las dos fórmulas y se opera.

$$R_{V_XV_Y} = \frac{r_{XY}}{\sqrt{r_{XX}}\sqrt{r_{YY}}} \qquad R_{V_XV_Y} = \frac{R_{XY}}{\sqrt{R_{XX}}\sqrt{R_{YY}}}$$

$$\frac{r_{XY}}{\sqrt{r_{XX}}\sqrt{r_{YY}}} = \frac{R_{XY}}{\sqrt{R_{XX}}\sqrt{R_{YY}}} \Rightarrow R_{XY} = \frac{r_{XY}}{\sqrt{\frac{r_{XX}}{R_{XX}}}\sqrt{\frac{r_{YY}}{R_{YY}}}}$$

Continuando con el ejemplo anterior, ¿cuál sería el coeficiente de validez si se consiguiera un coeficiente de fiabilidad en el test de 0,75 y en el criterio de 0,64?

$$R_{XY} = \frac{0,56}{\sqrt{\frac{0,64}{0,75}}\sqrt{\frac{0,60}{0,64}}} = \frac{0,56}{0,89} = 0,63$$

Se observa que ha habido un aumento del coeficiente de validez, ha pasado de 0,56 a 0,63. El aumento no es tan grande como cuando se consiguen eliminar por completo los errores de medida en el test y en el criterio, pero es bastante considerable.

4.2.5. Estimación del coeficiente de validez del test en el supuesto de que se mejorara la fiabilidad del test

Si se mejora la fiabilidad del test pero se mantiene constante la del criterio, la fórmula a utilizar sería:

$$R_{XY} = \frac{r_{XY}}{\sqrt{\frac{r_{XX}}{R_{XX}}}} \qquad [7.24]$$

El segundo radical del denominador desaparece ya que al permanecer constante el coeficiente de fiabilidad del criterio el numerador y el denominador son iguales.

¿Cuál sería el coeficiente de validez si el coeficiente de fiabilidad del test se pudiera aumentar hasta 0,75 y se mantuviera constante el del criterio?

$$R_{XY} = \frac{0,56}{\sqrt{\frac{0,64}{0,75}}} = \frac{0,56}{0,92} = 0,61$$

4.2.6. Estimación del coeficiente de validez del test en el supuesto de que se mejorara la fiabilidad del criterio

Siguiendo el mismo razonamiento anterior, la fórmula a utilizar sería:

$$R_{XY} = \frac{r_{XY}}{\sqrt{\frac{r_{YY}}{R_{YY}}}} \qquad [7.25]$$

Si se mantiene invariante el coeficiente de fiabilidad del test, y por algún procedimiento se consigue que la fiabilidad del criterio aumente hasta 0,64 ¿cuál sería el coeficiente de validez estimado?

$$R_{XY} = \frac{0,56}{\sqrt{\frac{0,60}{0,64}}} = \frac{0,56}{0,97} = 0,58$$

Aunque hay un aumento éste es bastante más pequeño.

4.2.7. Valor máximo del coeficiente de validez

Se obtiene a partir de la fórmula:

$$R_{V_X V_Y} = \frac{r_{XY}}{\sqrt{r_{XX}}\sqrt{r_{YY}}} \leq 1$$

que es la que permite estimar el coeficiente de validez cuando se han eliminado por completo los errores de medida del test y del criterio. Como toda correlación, es igual o menor que la unidad. Suponiendo que fuera igual a la unidad que sería el valor máximo, se deduce que:

$$r_{XY} \leq \sqrt{r_{XX}}\sqrt{r_{YY}}$$

y partiendo de la base de que el valor máximo del coeficiente de fiabilidad del criterio es la unidad, la fórmula anterior se podría expresar como:

$$r_{XY} \leq \sqrt{r_{XX}}$$

teniendo en cuenta que la raíz cuadrada del coeficiente de fiabilidad es el índice de fiabilidad, la fórmula anterior se puede expresar como:

$$r_{XY} \leq r_{VX} \qquad [7.26]$$

que indica que el valor máximo que puede alcanzar el coeficiente de validez es el del índice de fiabilidad.

En nuestro ejemplo, el valor máximo que podría alcanzar el coeficiente de validez empírico sería igual a 0,80.

4.3. Validez y longitud

Cuando se estudió el tema relativo a la fiabilidad de las puntuaciones se explicó que una de las formas de aumentar el coeficiente de fiabilidad era aumentando la longitud del test a base de añadirle elementos paralelos a los que ya tenía. Esta mejora del coeficiente de fiabilidad repercute, directamente, en una mejora del coeficiente de validez tal y como hemos expuesto en el apartado anterior; la relación entre el coeficiente de validez con la fiabilidad y la longitud del test viene dada por la siguiente expresión:

$$R_{XY} = \frac{r_{XY}\sqrt{n}}{\sqrt{1+(n-1)r_{XX}}} \qquad [7.27]$$

donde:

R_{XY} = coeficiente de validez estimado al modificar la longitud del test

r_{XY} = coeficiente de validez inicial del test, antes de la modificación de su longitud

n = número de veces que se aumenta o disminuye la longitud del test

r_{XX} = coeficiente de fiabilidad inicial del test

Para la deducción de la fórmula basta aplicar la ecuación general de Sperman-Brown que relaciona la fiabilidad y longitud:

$$R_{XX} = \frac{nr_{XX}}{1+(n-1)r_{XX}}$$

y sustituirla en la fórmula que relaciona la validez y la fiabilidad cuando se mejora el coeficiente de fiabilidad del test:

$$R_{XY} = \frac{r_{XY}}{\sqrt{\frac{r_{XX}}{R_{XX}}}} = \frac{r_{XY}}{\sqrt{\frac{r_{XX}}{\frac{nr_{XX}}{1+(n-1)r_{XX}}}}} = \frac{r_{XY}}{\sqrt{\frac{1+(n-1)r_{XX}}{n}}} = \frac{r_{XY}\sqrt{n}}{\sqrt{1+(n-1)r_{XX}}} \qquad [7.28]$$

Hay ocasiones que lo que se pretende es averiguar el número de veces que hay que aumentar o disminuir la longitud del test para conseguir un determinado valor del coeficiente de fiabilidad. En este caso, basta despejar el valor de n en la fórmula:

$$n = \frac{R_{XY}^2(1-r_{XX})}{r_{XY}^2 - R_{XY}^2 r_{XX}} \qquad [7.29]$$

Nota: Hay que tener en cuenta que n no es el número de ítems o elementos del test que hay que aumentar o eliminar para obtener un determinado coeficiente de fiabilidad o validez, n es el número de veces que hay que aumentar o disminuir la longitud del test y equivale, por lo tanto, al cociente entre el número de elementos finales y el número de elementos iniciales del test.

EJEMPLO:

Supongamos un test de 25 elementos que tiene un coeficiente de validez de 0,60 y un coeficiente de fiabilidad de 0,64. ¿Cuántos elementos habría que añadirle para obtener un coeficiente de validez de 0,70?

$$n = \frac{0{,}70^2(1-0{,}64)}{0{,}60^2 - 0{,}70^2 0{,}64} = \frac{0{,}1764}{0{,}0464} = 3{,}80$$

Este valor encontrado no indica que haya que añadir 3,8 ítems al test, lo que indica es que hay que aumentar la longitud del test en 3,8 veces; hay que hacerle 3,8 veces más largo. Para saber cuantos ítems representa ese aumento deberemos aplicar la siguiente fórmula:

$$n = \frac{EF}{EI} \Rightarrow EF = n \cdot EI = (3,8)(25) = 95$$

El test deberá tener 95 ítems para alcanzar un coeficiente de validez de 0,70. Habrá que añadirle, por lo tanto: 95 – 25 = 70 ítems.

Hay veces que el aumento en el coeficiente de validez no compensa el esfuerzo de añadir tantos elementos paralelos a un test; por otra parte, cuando un test se hace excesivamente largo se pueden introducir una serie de factores, como pueden ser el cansancio y la falta de motivación de los sujetos, que pueden aumentar los errores de medida.

5. GENERALIZACIÓN DE LA VALIDEZ

Ya hemos visto cómo la evolución del concepto de validez ha ido cambiando el énfasis de los aspectos externos de la misma a los internos. En la actualidad, uno de los enfoques más importantes

es la tendencia a la modelización de los procesos subyacentes a las respuestas a los ítems (Lachman, Lachman y Butterfield, 1979; Snow, Federico y Montague, 1980). Este cambio de enfoque en el estudio de la validez se refleja, como señalan Jones y Appelbaum (1989), en la conferencia sobre *Test Validity for the 1900's and Beyond* organizada en 1986 por *The Air Force Human resources and the Educational Testing Service* cuyas ponencias fueron posteriormente publicadas en un libro (Wainer y Braun, 1988). Tres capítulos del libro (Cronbach, 1988; Angoff, 1988 y Messick, 1988) hacen referencia a la teoría clásica de la validez, cada uno de ellos desde una perspectiva diferente, pero todos ellos subrayan la importancia de la validez de constructo sobre las demás.

Otra cuestión importante en los estudios de validez, que ha suscitado un gran interés en los últimos años, es el nivel de generalización de la misma. Mientras que este tema sólo era abordado de una manera superficial en la primera edición de *Standards for Educational and Psychological tests* (APA, AERA, NCME 1974) en la edición de 1985 se le dedica una atención especial y en 1986 el tratamiento había recibido tanta atención que se incluyó como una sección especial en el *Annual Review*.

El problema hace referencia a la posibilidad de utilizar y aplicar la evidencia obtenida en una situación a otras similares. Este problema reviste una enorme importancia, sobre todo en estudios de evaluación a gran escala, teniendo en cuenta que los estudios de validez suelen basarse en muestras de pequeño tamaño.

Desde 1986 se han hecho muchos estudios en este campo, las estrategias utilizadas son variaciones de los métodos tradicionales del meta-análisis (Glass, McGaw y Schmidt, 1981), lo que supone la reducción de los diversos resultados (codificados en función de sus características sustantivas y metodológicas) a una métrica común que haga factible su comparación y/o combinación. Las dos medidas que se suelen utilizar en el meta-análisis para transformar los resultados a una métrica común son los niveles de significación y el tamaño del efecto (coeficiente de correlación). Algunas modificaciones de este tipo de análisis han sido propuestas por Hunter, Schmidt y Coggin (1986).

Para una descripción clara de los procedimientos del meta-análisis véase Gómez-Benito (1987).

Otra aproximación al estudio de la posibilidad de generalización de la validez es la descrita por Hedhes (1988), que está basada en un método bayesiano de meta-análisis.

6. EJERCICIOS DE AUTOEVALUACIÓN

1. Un grupo normativo de 100 sujetos alcanzó una puntuación media de 15 puntos y una desviación típica de 5 en un test cuya fiabilidad era 0,91. Las calificaciones asignadas en un criterio arrojaron una media y una desviación típica de 10 y 4 puntos respectivamente. La fiabilidad del criterio fue 0,75 y la correlación entre las puntuaciones del test y las del criterio 0,80. Utilizando un N.C. del 95%, averiguar:

1.1. El tanto por ciento de la varianza del criterio que se debe al error.

1.2. El coeficiente de validez que se obtendría si se pudieran eliminar los errores de medida del test.

1.3. Entre qué valores se encontrará la puntuación en el criterio de un sujeto que en el test obtuvo 18 puntos.

2. Si el 19% de la varianza de las puntuaciones obtenidas en un test es varianza errónea y la correlación entre las puntuaciones verdaderas del test y las puntuaciones empíricas obtenidas en un criterio fuera de 0,85. ¿Cuál sería el coeficiente de validez empírico?

3. Si la incertidumbre con la que se puede pronosticar un criterio a partir de un test es del 60%. ¿ Cuál es el coeficiente de validez del test?.

4. Cierta Escuela de Enseñanza Superior desea cubrir las plazas de alumnos que quedan libres en el primer curso seleccionando los mejores de entre todos los aspirantes. Para llevar a cabo la selección se dispone de un test cuya correlación con el criterio de eficacia y éxito en la Escuela es de 0,75. El punto crítico en el criterio de éxito se ha situado en la media. Para ser admitido, la Escuela pide que se le garantice que el 90% de los elegidos va a tener un rendimiento aceptable. La media y la desviación típica del test utilizado fueron 16 y 5 y la del criterio 10 y 2 puntos respectivamente. La distribución de las puntuaciones en el test se ajusta a una distribución normal.

4.1. ¿Cuál sería la nota mínima que un sujeto debe obtener en el test para poder ser admitido? Expresar el resultado en puntuaciones directas.

4.2. Si un sujeto obtiene en el test una puntuación directa de 9 puntos, ¿cuál es la probabilidad de que fracase posteriormente en la Escuela?

5. A una oferta de trabajo publicada en un periódico (19-5-2002) se han presentado 400 licenciados universitarios de los que fueron admitidos los 20 que tuvieron mejores puntuaciones en un test utilizado para la selección. Las puntuaciones de los aspirantes en el test se distribuyeron según la curva normal de probabilidad con una media de 60 y una desviación típica de 4.

5.1. ¿Cuál fue la razón de selección?

5.2. ¿Cuál es la puntuación directa que como mínimo deben haber obtenido en el test los seleccionados?

6. Las puntuaciones directas obtenidas por un grupo de 100 sujetos en un test de rendimiento tienen una media de 11, una desviación típica de 1,20 y una fiabilidad de 0,91.

Examinados por un tribunal, la media de las calificaciones asignadas a los sujetos fue de 53 puntos y la desviación típica de 6. La fiabilidad del criterio era de 0,64 y la correlación entre las puntuaciones obtenidas en el test y las calificaciones asignadas por el tribunal 0,60.

Utilizando un N.C. del 95%, averiguar:

6.1. El error máximo de medida del test que se puede admitir a ese nivel de confianza.

6.2. ¿Cuál sería la verdadera correlación entre el test y el criterio si se eliminaran de éste todos los errores de medida que perturban su precisión?

6.3. Un sujeto obtuvo en el test una puntuación directa de 13,4 puntos. ¿Cuál es el intervalo confidencial en el cual podemos afirmar que estará comprendida su puntuación directa en el criterio?

6.4. Suponiendo que además del test de rendimiento se les hubiera aplicado a los sujetos un test de actitudes cuya correlación con el test de rendimiento fuera 0,49 y con el criterio de 0,54.

6.4.1. ¿Qué puntuación típica le pronosticaríamos en el criterio a un sujeto que obtuvo en el test de rendimiento una puntuación directa de 14 puntos y en el test de actitudes estuvo a una desviación típica por encima de la media?

6.4.2. ¿Entre qué valores estará la puntuación típica en el criterio de un sujeto que en el test de rendimiento estuvo a una desviación típica por debajo de la media y en el de actitudes se encontró en la media?

6.4.3. ¿Qué porcentaje de la varianza de las puntuaciones de los sujetos en el criterio se puede explicar a partir de los dos tests predictores?

6.4.4. Calcular el coeficiente de alienación y de valor predictivo múltiple y explicar los resultados obtenidos.

7. En una residencia de ancianos se está probando la validez de una escala de observación para detectar la dependencia funcional de los residentes y asignarles a un grupo de rehabilitación. A continuación se ofrecen las puntuaciones obtenidas por 11 residentes en la escala de observación y el diagnóstico emitido por los especialistas de la residencia en cuanto a su necesidad o no de rehabilitación.

Si se considerara que todos aquellos residentes que hubieran obtenido 20 puntos o más en la escala necesitaran rehabilitación:

7.1. ¿Cuál sería la validez predictiva de la escala?

7.2. ¿Qué punto de corte maximizaría las clasificaciones correctas?, asumiendo que en este caso la rehabilitación no perjudicaría a los residentes.

Sujetos	Escala	Diagnóstico
1	26	NR
2	11	NR
3	10	NR
4	6	NR
5	21	NR
6	25	R
7	18	R
8	15	NR
9	12	NR
10	30	R
11	32	R

8. Ejercicios conceptuales

Ante cada una de las afirmaciones que se muestran a continuación, el lector deberá responder si el concepto que contiene es verdadero o falso y justificar su respuesta.

1. La correlación múltiple es la correlación entre las puntuaciones obtenidas por los sujetos en una variable criterio y una variable predictora de la que se ha eliminado el efecto que pueda estar ejerciendo un conjunto de variables.
2. La correlación semiparcial es la correlación entre el criterio y una de las variables predictoras eliminando el efecto que sobre una de ellas puedan estar ejerciendo el resto de las variables.
3. La correlación parcial es la correlación entre el criterio y una de las variables predictoras cuando de dicha correlación se elimina el efecto que puedan estar ejerciendo el resto de las variables.
4. La correlación múltiple al cuadrado, multiplicada por ciento, representa el porcentaje de varianza errónea que hay en la varianza de las puntuaciones de los sujetos en el criterio.
5. La desviación típica de los errores de estimación es el error típico de estimación.
6. En el método fordward, que se utiliza para la selección de predictores, se calcula la correlación múltiple entre el criterio y el conjunto de variables predictoras de las que se dispone y, una a una, se van eliminando las variables que menos contribuyen a la medida del criterio.

7. El coeficiente Kappa permite evaluar la consistencia o acuerdo entre los decisores respecto a las decisiones adoptadas.
8. La sensibilidad es un índice de validez de las decisiones que equivale a la proporción de aspirantes que fueron correctamente rechazados en una selección.
9. A medida que aumenta la variabilidad de la muestra disminuye el coeficiente de validez.
10. El coeficiente de validez de un test puede aumentar si se le añaden elementos paralelos a los que ya tenía.

7. SOLUCIONES A LOS EJERCICIOS DE AUTOEVALUACIÓN

1. $N = 100$ Sujetos

$\overline{X} = 15 \quad S_X = 5 \quad r_{XX} = 0,91$

$\overline{Y} = 10 \quad S_Y = 4 \quad r_{YY} = 0,75 \quad r_{XY} = 0,80$

1.1. $r_{XY}^2 = 0,80^2 = 0,64$

El 64% de la varianza de las puntuaciones en el criterio se puede explicar a partir de la variable predictora, el complemento hasta el 100%, es decir un 36% es el porcentaje que queda sin explicar y, por lo tanto es el porcentaje correspondiente a la varianza residual o varianza error.

1.2.

$$R_{V_X Y} = \frac{r_{XY}}{\sqrt{r_{XX}}} = \frac{0,80}{\sqrt{0,91}} = 0,84$$

En el caso hipotético de que se pudieran eliminar todos los errores de medida del test, el coeficiente de validez aumentaría de 0,80 a 0,84.

1.3.

N.C. $95\% \Rightarrow Z_C = 1,96$

$$S_{Y \cdot X} = S_Y \sqrt{1 - r_{XY}^2} = 4\sqrt{1 - 0,80^2} = 4(0,6) = 2,4$$

$$E_{máx} = 2,4(1,96) = 4,70$$

$$Y' = r_{XY} \frac{S_Y}{S_X}\left(X - \overline{X}\right) + \overline{Y} = 0,80 \frac{4}{5}(18 - 15) + 10 = 11,92$$

$$Y' \pm E_{máx} = 11,92 \pm 4,70$$

$$7,22 \leq Y \leq 16,62$$

Se estima que la puntuación en el criterio de un sujeto que en el test haya obtenido una puntuación igual a 18 estará comprendida entre 7,22 y 16,62 con un nivel de confianza del 95%, o lo que es lo mismo con una probabilidad igual o menor de 0,05 de error.

2.

$$\frac{S_e^2}{S_X^2} = 0,19 \qquad r_{XX} = 1 - 0,19 = 0,81 \qquad R_{V_X Y} = 0,85$$

$$0,85 = \frac{r_{XY}}{\sqrt{r_{XX}}} = \frac{r_{XY}}{\sqrt{0,81}} \Rightarrow r_{XY} = 0,85\ (0,90) = 0,765 \simeq 0,77$$

El coeficiente de validez empírico es el real, el que se obtiene a partir de unos datos. Dado que no se pueden eliminar por completo los errores de medida ni del test ni del criterio, los coeficientes de validez obtenidos como correlación entre puntuaciones verdaderas y empíricas o entre dos series de puntuaciones verdaderas son coeficientes de validez teóricos. El valor del coeficiente de validez es 0,765.

3. La incertidumbre o inseguridad con la que se puede pronosticar un criterio a partir de un test viene dada por el coeficiente de alineación.

$$K = 0,60 = \sqrt{1 - r_{XY}^2} \Rightarrow 0,36 = 1 - r_{XY}^2 \Rightarrow r_{XY}^2 = 1 - 0,36 = 0,64$$

$$r_{XY} = 0,80$$

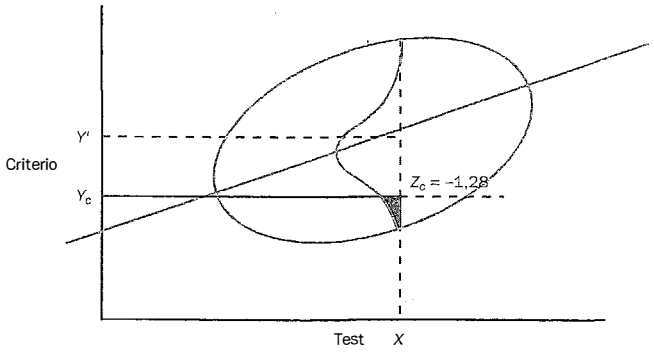
4.

$r_{XY} = 0,75$ $\quad \overline{X} = 16$ $\quad S_X = 5$

$\overline{Y} = 10$ $\quad S_Y = 2$ $\quad Y_C = 10$

4.1. Al punto de corte en el criterio, le corresponde una puntuación típica $Z_C = -1,28$ que es la que garantiza que hay una probabilidad de éxito del 90% de los elegidos mediante el test y cuya puntuación mínima debemos calcular.

$$S_{Y \cdot X} = S_Y \sqrt{1 - r_{XY}^2} = 2\sqrt{1 - 0,75^2} = 1,32$$



La puntuación típica correspondiente al 90% de éxito es igual a –1,28; es decir, es una puntuación típica que deja por debajo el 10% de la distribución de los errores de estimación.

$$Z_C = \frac{Y_C - Y'}{S_{Y \cdot X}} = \frac{10 - Y'}{1,32} = -1,28 \Rightarrow Y' = (1,28)(1,32) + 10 = 11,69$$

$$Y' = r_{XY} \frac{S_Y}{S_X}(X - \overline{X}) + \overline{Y} \Rightarrow\ 11,69 = 0,75 \frac{2}{5}(X - 16) + 10 = 0,3X + 5,2$$

Despejando $X \Rightarrow X = \dfrac{11,69 - 5,2}{0,3} = 21,63$

Esta sería la puntuación mínima que tendrían que obtener los sujetos en el test para que puedan ser admitidos con las garantías exigidas por la Escuela.

4.2. En primer lugar hay que conocer su puntuación pronosticada en el criterio:

$$Y' = 0,75 \frac{2}{5}(9 - 16) + 10 = 7,9$$

Después se calcula la puntuación típica que le corresponde en la distribución de los errores:

$$Z = \frac{10 - 7,9}{1,32} = 1,59$$

Se acude a las tablas de curva normal y se busca la probabilidad que deja por debajo una puntuación típica de 1,59. La probabilidad encontrada es 0,9441. Esta es la probabilidad de fracasar que tiene un sujeto que en el test obtuvo una puntuación de 9 puntos. Su probabilidad de éxito vendrá dada por la probabilidad que hay por encima de la puntuación típica $Z = 1,59$. Esta probabilidad es $1 - 0,9441 = 0,0559$.

5.

$N = 400 \qquad \overline{X} = 60 \qquad S_X = 4$

5.1. La razón de selección viene dada por la proporción de sujetos que han sido seleccionados.

$$R.S. = \frac{20}{400} = 0,05$$

Sólo han sido seleccionados el 5% de los aspirantes.

5.2. Los 20 seleccionados representan el 5% mejor de la muestra. Para ver cual ha sido la puntuación mínima que han obtenido en el test se busca la puntuación típica que deja por encima ese 5%. Esa puntuación típica es 1,64. A continuación, utilizando la fórmula de las puntuaciones típicas, se obtiene el valor de X.

$$1,64 = \frac{X - \overline{X}}{S_X} = \frac{X - 60}{4} \Rightarrow X = 1,64(4) + 60 = 66,56$$

Esa es la puntuación mínima que han tenido que obtener los sujetos.

6.

$N = 100 \qquad \overline{X} = 11 \qquad S_X = 1,20 \qquad r_{XX} = 0,\ 91$

$\overline{Y} = 53 \qquad S_Y = 6 \qquad r_{YY} = 0,64 \qquad r_{XY} = 0,60 \qquad NC\ 95\% \Rightarrow Z_C = 1,96$

6.1. El error máximo

$$S_e = S_X\sqrt{1 - r_{XX}} = 1,20\sqrt{1 - 0,91} = 0,36$$

$$E_{máx} = (1,96)\ (0,36) = 0,7056 \simeq 0,71$$

6.2.

$$R_{XV_Y} = \frac{r_{XY}}{\sqrt{r_{YY}}} = \frac{0,60}{\sqrt{0,64}} = 0,75$$

6.3.

$X = 13,4$

$$Y' = 0,60\frac{6}{1,20}(13,4 - 11) + 53 = 60,2$$

$$S_{Y\cdot X} = 6\sqrt{1 - 0,60^2} = 4,8$$

$E_{máx} = 1,96(4,8) = 9,41$

$I.C. = 60,2 \pm 9,41 \Rightarrow 50,79 \leq Y \leq 69,61$

La puntuación en el criterio estará entre los límites marcados por las puntuaciones 50,79 y 69,61 con una probabilidad del 95%.

6.4.

$TR = (X_1) \qquad TA = (X_2) \qquad \text{Criterio} = Y$

$r_{X_1X_2} = 0,49 \qquad r_{X_2Y} = 0,54 \qquad r_{X_1Y} = 0,60$

6.4.1. En primer lugar hay que construir la ecuación de regresión

$$Z'_Y = b_1^* Z_{X_1} + b_2^* Z_{X_2} = 0,44Z_{X_1} + 0,32Z_{X_2}$$

$$b_1^* = \frac{0,60 - 0,54\ .\ 0,49}{1 - 0,49^2} = \frac{0,3354}{0,7599} = 0,44$$

$$b_2^* = \frac{0,54 - 0,60\ .\ 0,49}{1 - 0,49^2} = \frac{0,246}{0,7599} = 0,32$$

Una vez construida la ecuación, se calculan las puntuaciones típicas correspondientes a las dos variables predictoras y, sustituyendo en la ecuación, se obtiene la puntuación pronosticada.

$$Z_{X_1} = \frac{14-11}{1,20} = 2,5$$

$$Z_{X_2} = 1$$

$$Z'_Y = 0,44(2,5) + 0,32(1) = 1,42$$

6.4.2. NC 95 % $Z_C = 1,96$

Para calcular el error típico de estimación en puntuaciones típicas es necesario conocer la correlación múltiple.

$$R^2_{Y \cdot X_1 X_2} = b^*_1 r_{YX_1} + b^*_2 r_{YX_2} = 0,44(0,60) + 0,32(0,54) = 0,44$$

$$R_{Y \cdot X_1 X_2} = \sqrt{0,44} = 0,66$$

El error típico de estimación múltiple en puntuaciones típicas será:

$$S_{Y \cdot X_1 X_2} = \sqrt{1-0,44} = 0,75$$

El error máximo = 0,75 · 1,96 = 1,47

Dado que la media de las puntuaciones típicas es cero, la puntuación pronosticada del sujeto en el criterio será:

$$Z'_Y = 0,44(-1) + 0,32(0) = -0,44$$

El intervalo confidencial:

$$-0,44 \pm 1,47 \Rightarrow -1,91 \leq Z_Y \leq 1,03$$

6.4.3.

$$R^2_{Y \cdot X_1 X_2} = 0,44$$

El 44% de la varianza de las puntuaciones de los sujetos en el criterio se puede explicar a partir de las dos variables predictoras; es decir, de los dos tests.

6.4.4.

$$K = \sqrt{1-R^2_{Y \cdot X_1 X_2}} = \sqrt{1-0,44} = 0,75$$

$$C.V.P. = 1-K = 1-0,75 = 0,25$$

El coeficiente de alineación K multiplicado por cien indica que al hacer los pronósticos hay un 75% de inseguridad o azar. Elevado al cuadrado nos informa del porcentaje de la varianza del criterio que no se puede explicar a partir de las variables predictoras, en nuestro caso un 56 %.

El coeficiente de valor predictivo es el complementario del coeficiente de alienación y multiplicado por cien indica el porcentaje de seguridad en los pronósticos, en nuestro caso un 25%.

7. En primer lugar se hace la tabla de doble entrada para ver como se distribuyen los sujetos en el test y en el criterio.

		Diagnóstico		
		R	NR	
Escala	R	3 (1,82)	2 (3,182)	5
	NR	1 (2,18)	5 (3,818)	6
		4	7	11

Hay 3 residentes a los que se les ha detectado la necesidad de rehabilitación tanto por el test como por el grupo de especialistas, 2 a los que se les detecta la necesidad de rehabilitación mediante el test pero los especialistas consideran que no la necesitan (falsos positivos), hay 1 que no es detectado por el test y sin embargo los especialistas consideran que si necesita rehabilitación (falso negativo) y, finalmente, en 5 residentes se ha considerado que no necesitan rehabilitación tanto a través del test como en opinión de los especialistas.

7.1. Para ver la validez predictiva de la escala se puede utilizar el coeficiente Kappa.

En la tabla aparecen entre paréntesis las frecuencias esperadas por azar que se han averiguado multiplicando las frecuencias marginales correspondientes y dividiendo por el total de sujetos.

$$K = \frac{8 - 5,64}{11 - 5,64} = \frac{2,36}{5,36} = 0,44$$

La escala tiene una validez media.

La proporción de clasificaciones correctas es: $\frac{8}{11} = 0,73$

El índice de sensibilidad: $\frac{3}{4} = 0,75$

La especificidad: $\frac{5}{7} = 0,71$

7.2. Para ver que punto de corte maximizaría las clasificaciones correctas vamos a ir probando con la puntuación 22 y con la 17. Ya hemos visto que cuando se toma como punto de corte para enviar a los residentes a rehabilitación una puntuación igual o mayor que 20, se comete 1 falso negativo y 2 falsos positivos.

Si se toma como punto de corte una puntuación igual a 22 se detectan 1 falso positivo y un falso negativo.

Si se toma como punto de corte una puntuación igual a 17 no se cometería ningún falso negativo y se cometerían 2 falsos positivos.

Ante estos resultados la decisión debería estar entre una puntuación igual a 22 o una puntuación igual a 17. Todo depende de las consecuencias de la decisión. Dado que la rehabilitación no perjudica a nadie, sería mejor poner el punto de corte en la puntuación igual a 17 pues de esta manera ningún residente que lo necesitara se quedaría sin rehabilitación (0 falsos negativos) y habría dos residentes que se beneficiarían de la rehabilitación sin necesitarla (2 falsos positivos) pero que no les vendría mal.

8. Respuestas a los ejercicios conceptuales

1. La afirmación es falsa.

   La correlación múltiple es la correlación entre las puntuaciones obtenidas por los sujetos en la variable criterio y las obtenidas en las variables predictoras tomadas conjuntamente.

2. La afirmación es verdadera.

3. La afirmación es verdadera.

4. La afirmación es falsa.

   La correlación múltiple al cuadrado (multiplicada por cien) expresa el porcentaje de varianza común o asociada entre el criterio y el conjunto de variables predictoras o, dicho de otro modo, el porcentaje de la variación de las puntuaciones de los sujetos en el criterio que se puede explicar a partir de la variación de las puntuaciones de los sujetos en el conjunto de variables predictoras.

5. La afirmación es correcta.

6. La afirmación es incorrecta.

   Ese sería el método backward, en el método forward se comienza incluyendo la variable que tiene una correlación más alta con el criterio y, paso a paso (una a una) se van incorporando a la ecuación de regresión las distintas variables en función de su correlación con el criterio.

7. La afirmación es correcta.

8. La afirmación es incorrecta.

   La sensibilidad es un índice de validez que representa la proporción de aspirantes correctamente seleccionados mediante la prueba o test respecto al total de los que obtuvieron éxito en el criterio.

9. La afirmación es incorrecta.

   El coeficiente de validez, como cualquier coeficiente de correlación, aumenta con la variabilidad de la muestra.

10. La afirmación es correcta.

    Al aumentar el número de elementos de un test a base de añadirle elementos paralelos a los que ya tenía, el coeficiente de fiabilidad aumenta. Dada la relación entre la validez y la fiabilidad de los tests, este aumento en el coeficiente de fiabilidad incide en un aumento en el coeficiente de validez. Sin embargo, el valor máximo del coeficiente de validez es el índice de fiabilidad; por ello, llega un momento en que por más que se aumente el número de ítems no se puede aumentar el coeficiente de validez.

8. BIBLIOGRAFÍA COMPLEMENTARIA

Martínez – Arias, R.; Hernández Lloreda, Mª J.; Hernández Lloreda, Mª V. (2006). *Psicometría*. Madrid: Alianza editorial

Martínez – Arias, M.R. (1995). *Psicometría: Teoría de los tests psicológicos y educativos*. Madrid: Síntesis. Capítulo 18

Muñiz, J. (1998). *Teoría Clásica de los Tests*. Madrid: Pirámide. Capítulo 4.

Navas, M.J. (2001). *Métodos, diseños y técnicas de investigación psicológica*. Madrid: UNED. Capítulo 7.

Santisteban, C. (1990). *Psicometría*. Madrid: Norma. Capítulo 15.

TEMA 8

ANÁLISIS DE LOS ÍTEMS

Francisco Pablo Holgado Tello

SUMARIO

1. ORIENTACIONES DIDÁCTICAS

El *análisis de los ítems* se puede definir como un proceso por el que los ítems de un test son evaluados y examinados críticamente con el objetivo de identificar y reducir las fuentes de error, tanto aleatorio como sistemático para poder eliminar aquellos que no reúnen las suficientes garantías psicométricas. Es frecuente que los constructores de tests lleven a cabo este tipo de análisis para obtener evidencias sobre la calidad de los ítems con el objetivo de identificar aquellos que se han de descartar del test final, reformular otros que puedan ser mejorados, y conservar los que definitivamente presenten unas buenas propiedades psicométricas (Osterlind, 1998).

El análisis de los ítems, al contrario de lo que pudiera parecer, comienza con el proceso de su redacción, proceso en el que hay que atender a toda una serie de directrices (consultar tema 2) antes de plantearse la administración de los mismos. Estas directrices facilitarán una adecuada construcción de los propios ítems (incluyendo sus alternativas) y hará más probable que obtengamos un test de mayor calidad, que se verá plasmada en las propiedades psicométricas del mismo (Shultz y Whitney, 2005). Tanto los enunciados de los ítems como sus alternativas, si están cuidadosamente redactadas, redundaran directamente en la obtención de buenos indicadores sobre la calidad con la que conjuntamente miden el constructo de interés (Martínez, Moreno y Muñiz, 2005).

Habitualmente, los tests están construidos a partir de un número determinado de elementos. En teoría, si conociéramos la calidad de cada uno de ellos podríamos deducir la calidad psicométrica del test. Es decir, el conocimiento de las características individuales de los ítems puede dar indicios para mejorar el test y maximizar sus propiedades psicométricas, lo que se traduce en una reducción del error aleatorio, con el consiguiente aumento de la fiabilidad, a la hora de medir aquellas conductas del domino de interés que pretende medir (validez).

Una estrategia general consiste en determinar el número de ítems necesario para confeccionar el test. Este aspecto se puede establecer a partir del tiempo total disponible, o del tiempo estimado en responder a cada ítem. Generalmente, con el objetivo de analizar la calidad métrica de los ítems

y obtener un test con una fiabilidad óptima mediante el menor número de ítems posible, un número concreto de ítems (2, 3, o hasta 4 veces mayor que el número de ítems del test resultante) se administra a una muestra piloto de sujetos con características similares a la población diana, o población a la que va dirigido el test. Hay que resaltar que al seleccionar los ítems, es conveniente tener en cuenta la finalidad del test, y las decisiones que se tomarán a partir de sus puntuaciones, ya que en función de ello se utilizarán ítems con niveles de dificultad diferentes (ver tema 2).

En líneas generales, el análisis de los elementos de un test se puede llevar a cabo mediante dos procedimientos que son complementarios y que ofrecen mucha información al constructor del test sobre el comportamiento métrico de los elementos que lo van a formar. Por un lado, se pueden realizar análisis numéricos y cuantitativos que tratan de obtener determinados estadísticos basados en la distribución de las respuestas de los sujetos a las distintas alternativas de los ítems y, por otro, se puede recurrir a juicios de expertos en el que se cuestiona la calidad métrica del ítem en función de criterios conocidos. Uno de los aspectos más relevantes que se evalúan mediante un juicio de expertos es la validez de contenido de los ítems, para lo que se han desarrollado diversos índices de acuerdo o congruencia entre jueces como por ejemplo, el CRV (*content validity ratio*) propuesto por Lawshe (1975) que se basa en un promedio que toma como referencia el número de jueces que consideran un ítem «no-necesario», «útil» o «esencial» para medir el dominio al que ha sido asignado. Sin embargo, uno de los más utilizados (Osterlind, 1998), es el índice de congruencia propuesto por Rovinelli y Hambleton (1977) y Hambleton (1980) en el que para evaluar la validez de contenido de cada ítem se le pide a cada uno de los jueces que valoren en una escala de tres puntos (–1; 0; 1), o de cinco según proponen recientemente Sanduvete, Chacón, Sánchez y Pérez-Gil (2013), el grado en que el ítem está relacionado con la dimensión que trata de medir. De manera ideal un buen análisis de los ítems ha de contemplar tanto el juicio de expertos sobre la calidad y adecuación de los ítems, como una serie de análisis numéricos que proporcionen distintos estadísticos; es decir, un buen análisis de ítems debe incluir tanto el análisis cualitativo como el análisis cuantitativo de los mismos.

Una vez que hemos analizado la calidad psicométrica de los elementos de un test, y disponemos de las garantías suficientes acerca de su adecuación, es cuando se analiza la calidad global del test, donde destacan dos conceptos fundamentales en Psicometría: fiabilidad y validez.

> **Nota:** El motivo de situar este tema justo después de los de fiabilidad y validez es por razones didácticas ya que sería complicado tratar algunas propiedades de los ítems sin haber explicado anteriormente ambos conceptos básicos en el proceso de medición en Psicología.

En este tema vamos a abordar el estudio del análisis cuantitativo o estadístico de los ítems, puesto que el análisis cualitativo se ha abordado ya a lo largo de los temas precedentes, y aunque son muchas las propiedades y características que podemos estudiar nos centraremos en aquellas que van a afectar a la calidad global del test (Muñiz, 2003): la dificultad de los ítems, su poder discriminativo, el análisis de los distractores o alternativas incorrectas de respuesta y la fiabilidad y validez de los ítems. Además, abordaremos una importante cuestión directamente relacionada con la validez como es el funcionamiento diferencial de los ítems.

En este tema, es necesario atender a los siguientes objetivos:

- Conocer las propiedades psicométricas de los ítems.
- Saber calcular los estadísticos que, desde la Teoría Clásica de los Tests, se han propuesto para evaluar la calidad métrica de los ítems.
- Reconocer la importancia que tiene el análisis de las alternativas incorrectas para la mejora de la calidad de los ítems. Y saber realizar un análisis de distractores.
- Conocer cómo se relacionan las propiedades psicométricas de los ítems con las del test total.
- Saber en qué consiste el concepto de Funcionamiento Diferencial de los Ítems (FDI) y saber cómo detectarlo.

2. INTRODUCCIÓN

Como hemos visto en temas anteriores, los ítems pueden adoptar distintos formatos y evaluar variables cognitivas (aptitudes, rendimiento, etc.) donde hay respuestas correctas e incorrectas, o variables no cognitivas (actitudes, intereses, valores, etc.) donde no hay respuestas correctas. Los estadísticos que presentamos se utilizan, fundamentalmente, con ítems aptitudinales o de rendimiento en los que existe una alternativa correcta y una o varias incorrectas.

Para llevar a cabo un análisis de ítems, en primer lugar se debe disponer de una matriz de datos con las respuestas de los sujetos a cada uno de los ítems. Tanto para el análisis de las puntuaciones del test como de las respuestas a la alternativa correcta, la matriz tomará la forma de unos y ceros, donde los unos hagan referencia a los aciertos, mientras que los ceros harán referencia a los fallos. Para el análisis de las alternativas incorrectas, en la matriz han de aparecer las opciones concretas que haya seleccionado cada sujeto.

El análisis de la alternativa correcta, que es la que ofrece más información sobre la calidad del test, permite obtener el índice de dificultad, el de discriminación y la fiabilidad y la validez del ítem. Muy brevemente diremos que la dificultad empírica de un ítem hace alusión a la proporción de sujetos que lo responden correctamente. Aunque la dificultad de un ítem puede establecerse teóricamente a priori en el proceso de redacción de acuerdo con la complejidad estimada del ítem, lógicamente, tendrá que ser contrastada con la dificultad empírica, que es la que presentamos en este tema. Hay veces que el constructor de la prueba piensa que un ítem tiene una dificultad pequeña y, a la hora de la verdad, resulta difícil y viceversa. El poder discriminativo indica la capacidad del ítem para diferenciar a los sujetos con distinto nivel en el rasgo medido. Ambos estadís-

ticos están directamente relacionados con la media y varianza de las puntuaciones totales del test. La fiabilidad y validez de los ítems están relacionadas con la desviación típica del test e indican la posible contribución de cada ítem a la fiabilidad y validez de las puntuaciones totales del test.

El análisis de las respuestas incorrectas o distractores aporta evidencias sobre la utilidad de cada alternativa y su contribución a la calidad métrica del ítem. Por tanto, su revisión es fundamental para mejorar el ítem en cuestión, mediante la sustitución o reparación de los distractores que no funcionen como tales.

Finalmente, un aspecto a evaluar dentro del análisis de ítems, es si de manera sistemática sujetos de distintos grupos de pertenencia pero con el mismo nivel en el rasgo medido tienen distintas probabilidades de éxito en el ítem en cuestión (Shultz y Whitney, 2005). A esta circunstancia se la conoce como funcionamiento diferencial de los ítems (FDI).

3. DIFICULTAD DE LOS ÍTEMS

Probablemente uno de los índices más populares para cuantificar la dificultad de los ítems, dicotómicos o dicotomizados, es la proporción de sujetos que han respondido correctamente al mismo. Hay que decir, que la dificultad así considerada es relativa, ya que va a depender del número de personas que intentan responder al ítem y de sus características, puesto que no se obtendrá el mismo índice de dificultad si el ítem dado es respondido por una muestra de superdotados que por otra de sujetos normales. Formalmente el índice de dificultad viene expresado por:

$$ID = \frac{A}{N} \qquad [8.1]$$

donde:

A = número de sujetos que aciertan el ítem.

N = número de personas que intentan responder al ítem.

El índice de dificultad oscila entre 0 y 1. Donde 0 indica que ningún sujeto ha acertado el ítem, y por lo tanto se trata de un ítem difícil, mientras que 1 hace referencia a que todos los sujetos respondieron correctamente el ítem indicando por tanto que se trata de un ítem fácil. Es por ello que, en realidad, debería llamarse índice de facilidad más que de dificultad, puesto que cuanto más próximo a 1 sea el *ID*, más fácil resulta el ítem. En general, se recomienda que los ítems con valores extremos para la población a la que van dirigidos sean eliminados del test final ya que no contribuyen a diferenciar entre sujetos con distinto nivel en el rasgo medido, puesto que o todos los aciertan o todos los fallan. Ahora será más fácil entender que si un ítem se aplica a una muestra de superdotados, su índice de dificultad será mucho mayor que si se administra a una muestra de sujetos normales pero, a la hora de interpretarlo, lógicamente, a los primeros les ha resultado mucho más fácil que a los segundos.

EJEMPLO:

Imaginemos que un ítem de rendimiento en matemáticas se aplica a 10 sujetos con el resultado mostrado en la siguiente tabla donde las letras hacen referencia a sujetos:

TABLA 8.1
Tipos de correlaciones en función del tipo de variables implicadas

Sujeto	A	B	C	D	E	F	G	H	I	J
Respuesta	1	1	1	1	0	1	0	1	1	0

Es decir, de los 10 sujetos que han intentado responder al ítem, 7 lo han acertado mientras que 3 han fallado. Ello se traduce en que el *ID* será de 0,70.

$$ID = \frac{7}{10} = 0{,}70$$

El valor de 0,70 obtenido no indica nada sobre si el ítem es bueno o malo. Simplemente representa cuánto de difícil ha resultado para la muestra de sujetos que lo han intentado responder. Si el mismo ítem fuera administrado a otra muestra de sujetos muy probablemente el *ID* sería distinto. Es decir, la dificultad es dependiente de la muestra de sujetos utilizada.

Concretamente, el dato proporcionado por el *ID* resulta de mucho interés en los Tests Referidos al Criterio (TRC), así si un grupo de ítems que miden el mismo concepto han resultado muy fáciles para un conjunto de alumnos podría pensarse que no tiene mucho sentido evaluar dicho dominio en esta muestra de sujetos dado que dominan el concepto medido. Por el contrario, si dicho grupo de ítems resultara muy difícil, entonces habría que pensar que la instrucción realizada no ha sido adecuada, por ejemplo.

El *ID* está relacionado directamente con la media y varianza del test. Respecto a la media, en ítems dicotómicos encontramos la siguiente relación:

$$ID = \frac{\sum_{j=1}^{n} X_j}{N} \qquad [8.2]$$

donde:

X_j = puede ser 1 o 0 según se acierte o falle el ítem.

Por tanto, para un ítem concreto llegamos fácilmente a la conclusión de que $\Sigma X_j = A$ (aciertos); es decir la suma de todas las puntuaciones obtenidas por los sujetos en ese ítem es igual al número de aciertos y por lo tanto el índice de dificultad del ítem es igual a su media. Si generalizamos al test total encontramos que la media de las puntuaciones en el test es igual a la suma de los índices de dificultad de los ítems (García-Cueto, 2005).

$$\overline{X} = \sum_{j=1}^{n} ID_j \qquad [8.3]$$

De esta forma imaginemos que el ítem anterior forma parte de un test compuesto por 5 ítems, tal y como se muestra en la siguiente tabla.

TABLA 8.2
Datos del ejemplo

Sujetos	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5	Total
A	1	1	1	1	1	5
B	1	0	1	0	1	3
C	1	1	0	1	0	3
D	1	0	0	1	0	2
E	0	1	0	1	1	3
F	1	0	0	1	0	2
G	0	1	1	1	0	3
H	1	0	0	1	0	2
I	1	1	0	1	0	3
J	0	0	0	1	1	2
ID	0,70	0,50	0,30	0,90	0,40	$\sum_{i=1}^{n} ID = 2,80$

$$\overline{X} = \frac{5+3+3+2+3+2+3+2+3+2}{10} = 2,80$$

$$\sum_{j=1}^{5} ID = 0,70 + 0,50 + 0,30 + 0,90 + 0,40 = 2,80$$

La relación entre la dificultad y la varianza del test es aún más directa, sobre todo si consideramos que en ítems dicotómicos la varianza viene dada por:

$$S_j^2 = p_j q_j \qquad [8.4]$$

donde:

p_j = proporción de sujetos que responden correctamente al ítem, es decir, el *ID*.

$q_j = 1 - p_j$.

Por tanto, la relación entre la dificultad del ítem y su varianza es directa. Dentro del análisis de los ítems, una cuestión muy relevante es encontrar el valor de p_j que maximiza la varianza del ítem. Observando la ecuación 8.4, se encontrará fácilmente una respuesta a esta cuestión, dado que la varianza máxima la alcanza un ítem cuando su p_j es igual a 0,5. Para llegar a esta conclusión basta con ir sustituyendo p_j por valores entre 0 y 1 y calcular la varianza.

Es lógico suponer que ítems acertados o fallados por todos los sujetos presentan una varianza igual a cero. Ello implica que no hay variabilidad en las respuestas, es decir, todas las respuestas son ceros o unos y por lo tanto cualquier sistema de clasificación basado en este ítem es inútil ya que siempre clasificaría a los sujetos en el mismo lugar. Un ítem es adecuado cuando al ser respondido por distintos sujetos provoca en ellos respuestas diferentes. Este aspecto está directamente relacionado con la *discriminación*, concepto que veremos más adelante.

3.1. Corrección de los aciertos por azar

En el cálculo del índice de dificultad hay que tener en cuenta que el hecho de acertar un ítem no sólo depende de que los sujetos conozcan la respuesta, sino también de la suerte que tengan aquellos que sin conocerla eligen la alternativa correcta. De esta forma cuanto mayor sea el número de distractores (o alternativas incorrectas) menos probable es que los sujetos acierten el ítem por azar puesto que habrá más alternativas para elegir.

Es decir, si en una muestra de sujetos hubiera algunos de ellos que no conociendo la respuesta a ningún ítem, **sistemáticamente respondieron a todos**, entonces acertarían un número determi-

nado de ítems por azar. Así por ejemplo, si un sujeto con una aptitud nula (o con un conocimiento nulo de la materia si se tratara de una prueba de conocimientos) respondiera a 25 ítems de 3 alternativas equiprobables acertaría por azar 1/3 de los mismos (aproximadamente 8). Lo que provoca que el número de aciertos sea mayor que los esperados en función del nivel de aptitud de los sujetos. Por ello, se aconseja corregir el *ID*. Tal y como se vio en el capítulo 2, es posible que los sujetos dejen ítems sin responder (omisiones), lo que implicaría otro tipo de corrección que vendría dada por los ítems que hubiera acertado si los hubiera respondido, aunque fuera por azar. No obstante, en este capítulo y para simplificar cálculos y conceptos vamos a considerar que no hay omisiones. En el caso de que en algún ejercicio las hubiera, se indicaría qué hacer con ellas.

$$ID_c = \frac{A}{N} - \frac{\frac{E}{k-1}}{N} = p - \frac{q}{k-1} \qquad [8.5]$$

donde:

ID_c = índice de dificultad corregido.

A = aciertos.

E = errores.

p = proporción de aciertos.

q = proporción de errores.

k = número de alternativas del ítem.

N = número de personas que intentan responder al ítem.

Así, si el test anterior estuviera compuesto por ítems de tres alternativas de respuesta, los índices de dificultad serían:

TABLA 8.3
Datos del ejemplo

Sujetos	Ítem 1	Ítem 2	Ítem 3	Ítem 4	Ítem 5
A	1	1	1	1	1
B	1	0	1	0	1
C	1	1	0	1	0
D	1	0	0	1	0
E	0	1	0	1	1
F	1	0	0	1	0
G	0	1	1	1	0
H	1	0	0	1	0
I	1	1	0	1	0
J	0	0	0	1	1
ID	0,70	0,50	0,30	0,90	0,40
ID_C	0,55	0,25	−0,05	0,85	0,10

$$ID_{c_1} = p - \frac{q}{k-1} = 0{,}70 - \frac{0{,}30}{2} = 0{,}55$$

.
.
.

$$ID_{c_5} = p - \frac{q}{k-1} = 0{,}40 - \frac{0{,}60}{2} = 0{,}10$$

Comparando las dos últimas filas de la tabla anterior se observa que los ítems que han sufrido una corrección mayor son los que han resultado más difíciles, como por ejemplo el 3. Se supone que habrá mayor número de aciertos por azar en los ítems más complicados, dado que los sujetos tienden a desconocer la respuesta correcta. De hecho, el ID_C del ítem 3 llega a ser negativo, lo que podría estar indicando que este ítem ha podido ser acertado más veces por azar que por conocer la respuesta correcta. En realidad cuando esto ocurre es que los sujetos han respondido al azar e incluso el azar les juega en su contra pues tienen más aciertos atribuidos al azar que los que realmente tienen porque saben de qué están hablando. Mientras que en los ítems fáciles los sujetos responderán, en mayor medida, desde el conocimiento del contenido del ítem, por lo que la corrección de aciertos por azar es más leve.

En la selección de los ítems que han de formar parte del test, la dificultad no es una cuestión baladí. Como norma general, en los tests de aptitudes se consiguen mejores resultados psicométricos cuando la mayoría de los ítems son de dificultad media. Lógicamente habrá que incluir ítems fáciles, situados preferentemente al comienzo del test para que el examinando no se desmotive, e ítems difíciles. Los primeros serán útiles para medir a los sujetos menos competentes, mientras que los segundos permitirán identificar al grupo de sujetos con mejor nivel en el rasgo medido por el test.

4. DISCRIMINACIÓN

Otro pilar fundamental en el análisis de los ítems responde al nombre de *discriminación*. La lógica que subyace a este concepto es que dado un ítem, los sujetos con buenas puntuaciones en el test han de acertarlo en mayor proporción que los que tienen bajas puntuaciones. El caso contrario, estaría indicando que precisamente los sujetos con más competencia tienden a fallar el ítem, mientras que los sujetos menos aptos lo aciertan en su mayoría, lo que va en contra del sentido común. Por otra parte si un ítem no sirve para diferenciar entre los sujetos en función de su nivel de competencia; es decir, no discriminara entre los sujetos, debería eliminarse.

Cuando se seleccionan ítems con poder discriminativo es porque se pretende diferenciar a los sujetos en función de su nivel en el rasgo medido. Una primera aproximación intuitiva al cálculo de la discriminación de un ítem implicaría contrastar la proporción de aciertos entre dos grupos extremos de aptitud, uno bajo y otro alto. Si el ítem discriminara adecuadamente, una consecuencia directa sería que la proporción de aciertos en el grupo de alta aptitud sería mayor que en el de baja aptitud; o lo que es lo mismo, que la correlación entre las puntuaciones obtenidas por los sujetos en el ítem y las obtenidas en el test total sería positiva. En base a ello, se han propuesto distintas formas para estudiar el poder discriminativo de los ítems.

4.1. Índice de discriminación basado en grupos extremos

El índice de discriminación *D* se basa en las proporciones de aciertos entre grupos extremos de aptitud. Kelly (1939) aconseja tomar el 27% (o el 25%) superior y el 27% (o el 25%) inferior de la muestra total para obtener un índice *D* sensible y estable. Es decir, el 27% superior estaría formado por los sujetos que han puntuado por encima del percentil 73 en el test total, mientras que el inferior por aquellos otros con puntuaciones por debajo del percentil 27. Una vez conformados los grupos se calcula la proporción de respuestas correctas a un determinado ítem en ambos grupos y se aplica la siguiente ecuación:

$$D = p_s - p_i \qquad [8.6]$$

donde:

p_s = proporción de aciertos en el grupo superior.

p_i = proporción de aciertos en el grupo inferior.

El índice *D* oscila entre –1 y 1. Tomará el valor ideal de 1 cuando todas las personas del grupo superior hayan acertado el ítem y las del inferior lo hayan fallado. Si *D* fuera igual a 0, estaría indicando que el ítem es acertado indistintamente en ambos grupos, es decir, estar en un grupo u otro es indiferente para acertar o no el ítem. *D* tomará valores negativos cuando los sujetos menos competentes acierten el ítem en mayor medida que los más competentes, lo que no es razonable porque dicho resultado estaría indicando que el ítem confunde a los más hábiles.

La discriminación también se puede representar gráficamente de forma que se vea claramente cómo puede interpretarse como la proporción de aciertos en función del nivel de aptitud de los sujetos. De esta forma, un ítem con un índice *D* alto quedaría representado tal y como aparece en la figura 8.1.

El ítem presentado en la figura 8.1 permite diferenciar a los sujetos en función de su nivel de aptitud. A medida que el nivel de habilidad de los sujetos se incrementa la probabilidad de acertar el ítem es mayor. Es decir, el grupo de alta aptitud lo acierta en mucha mayor proporción que los de baja aptitud.

FIGURA 8.1
Ítem con alto poder discriminativo

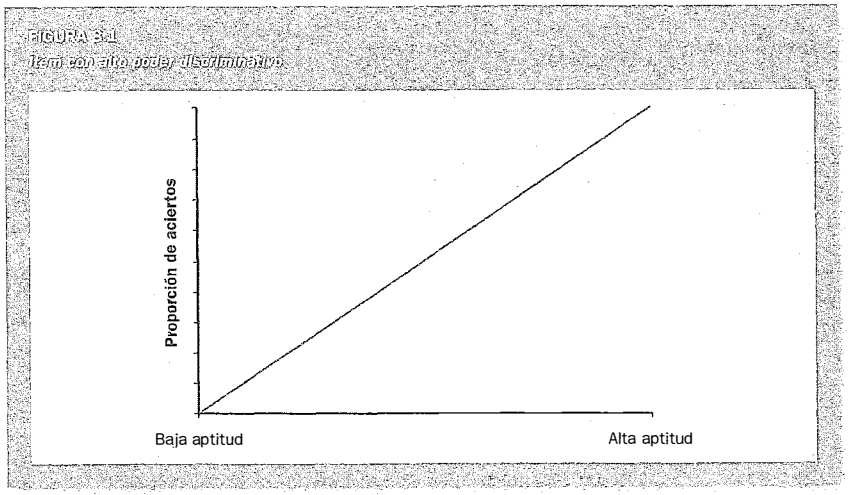
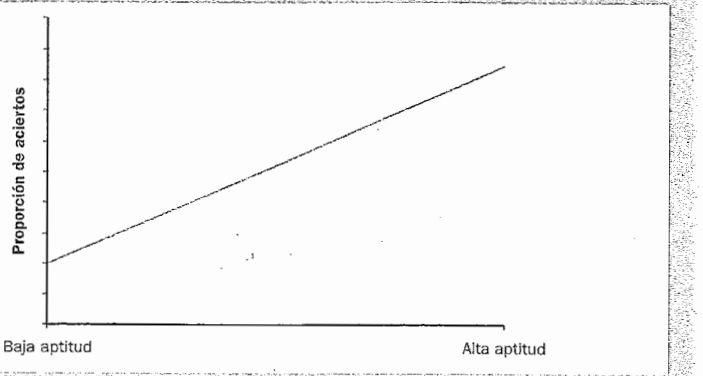
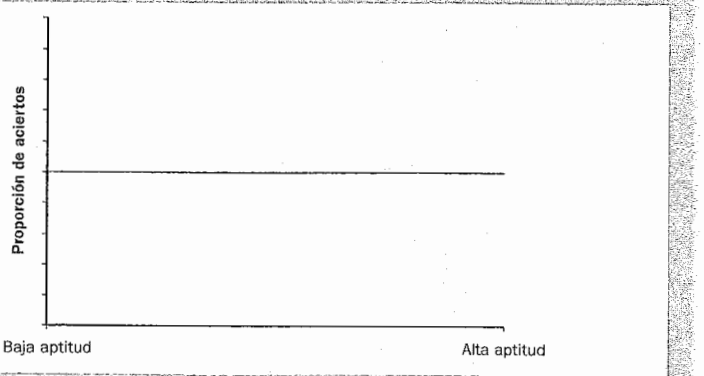


FIGURA 8.2
Ítem con poder discriminativo moderado



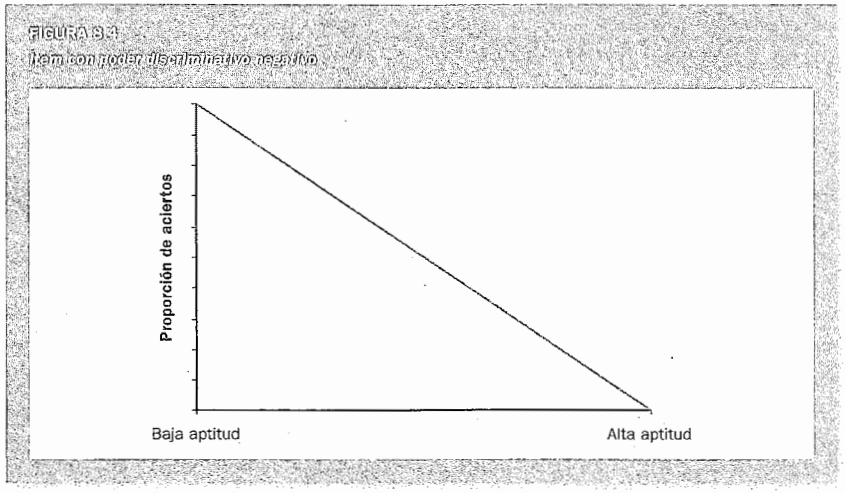
La figura 8.2 representa un ítem con una discriminación moderada. A pesar de que permite separar entre sujetos con distinto nivel de aptitud, no lo hace con toda la rotundidad que el representado en la figura 8.1 ya que hay un porcentaje de sujetos con baja aptitud que tienden a acertar el ítem (ver ordenada en el origen), y de entre los sujetos con alta aptitud existen otros tantos que tienden a fallarlo.

FIGURA 8.3
Ítem con poder discriminativo nulo



Observando la figura 8.3, se entiende fácilmente que se trata de un ítem que no representa ningún tipo de poder discriminativo. La proporción de aciertos no es función del nivel de aptitud de los sujetos, con lo que tampoco podríamos diferenciarlos en función de que hayan respondido correcta o incorrectamente a este ítem. El resultado es una línea horizontal, lo que indica que ambos grupos tienen la misma probabilidad de acertar el ítem.

FIGURA 8.4
Ítem con poder discriminativo negativo



Finalmente, en la figura 8.4 se presenta el caso de un ítem que discrimina en sentido contrario al que cabría esperar. Es decir, los sujetos con menos competencia tienden a acertarlo en mayor grado que los más hábiles, a los que probablemente está confundiendo por alguna razón que habría que investigar y corregir.

EJEMPLO:

En la tabla 8.4 aparecen las respuestas dadas por 370 sujetos a las 3 alternativas (A, B, C) de un ítem, donde la opción B es la correcta. Por filas aparece la frecuencia de sujetos que han seleccionado cada alternativa y que han obtenido puntuaciones superiores e inferiores al 27% de su muestra en el test total, así como el grupo conformado por el 46% central.

TABLA 8.4
Sujetos por nivel de aptitud

	A	B*	C
27% superior	19	53	28
46% intermedio	52	70	48
27% inferior	65	19	16

Calcular el índice de dificultad corrigiendo el efecto del azar, y el índice de discriminación.

La proporción de respuestas correctas será igual (53 + 70 + 19)/370 = 0,38; mientras que la proporción de errores será 228/370 = 0,62, luego el ID_c es igual a:

$$ID_c = p - \frac{q}{k-1} = 0,38 - \frac{0,62}{3-1} = 0,07$$

Para calcular *D* nos valemos exclusivamente de los grupos extremos:

$$D = p_s - p_i = \frac{53}{19+53+28} - \frac{19}{65+19+16} = \frac{53-19}{100} = 0,34$$

El marco de referencia para interpretar los valores de *D* lo proporciona Ebel (1965) mediante la siguiente tabla:

TABLA 8.5
Interpretación del nivel de discriminación

Valores	Interpretación
$D \geq 0,40$	El ítem discrimina muy bien
$0,30 \leq D \leq 0,39$	El ítem discrimina bien
$0,20 \leq D \leq 0,29$	El ítem discrimina poco
$0,10 \leq D \leq 0,19$	El ítem necesita revisión
$D < 0,10$	El ítem carece de utilidad

A la vista de los resultados el ítem resulta difícil, pero discrimina razonablemente bien.

4.2. Índices de discriminación basados en la correlación

Si un ítem discrimina adecuadamente entonces la correlación entre las puntuaciones obtenidas por los sujetos en el ítem y las obtenidas en el test total será positiva. Es decir, los sujetos que puntúan alto en el test tendrán más probabilidad de acertar el ítem. Este extremo, se puede observar en las figuras anteriores, de tal forma que si en un eje colocamos la puntuación en el test y en otro la puntuación en el ítem, un ítem con una discriminación adecuada presentará una correlación positiva (figura 8.1 y 8.2); si la discriminación fuera nula la correlación sería igual a cero, lo que se corresponde con la figura 8.3; y si discriminara en sentido inverso su correlación sería negativa (figura 8.4). Por tanto, podríamos definir la discriminación como la correlación entre las puntuaciones de los sujetos en el ítem y sus puntuaciones en el test (Muñiz, 2003). Lógicamente, la puntuación total de los sujetos en el test ha de calcularse descontando la puntuación del ítem. En caso contrario, estaríamos incrementando artificialmente el índice de discriminación ya que estaríamos correlacionando una variable (ítem) con otra variable (test) que contiene a la primera.

En el párrafo anterior implícitamente se ha hecho referencia a un concepto muy importante en Psicometría y que también fue tratado en el apartado sobre Likert del tema 3. Concretamente nos referimos a la relación que existe entre la probabilidad de acertar un ítem con el nivel de aptitud o rasgo medido. A este concepto se le denomina Curva Característica del Ítem (CCI) y es importante porque es posible modelar dicha relación matemáticamente a partir de los parámetros de dificultad, discriminación y acierto por azar. Sin embargo, no profundizaremos en estos aspectos porque exceden ampliamente los objetivos de este tema.

Ahora bien, el índice de correlación utilizado ha de ser coherente con el tipo de puntuaciones del ítem y del test. En el tema 6 sobre Validez, ya se expusieron los tipos de correlación adecuados para cada tipo de variable.

Los coeficientes que veremos a continuación son la correlación *Phi* (Φ), la biserial-puntual y la biserial.

4.2.1. Coeficiente de correlación Φ

Se utiliza cuando las puntuaciones del ítem y del test son estrictamente dicotómicas. Su principal utilidad reside en que permite estimar la discriminación de un ítem con algún criterio de interés. De esta forma, podríamos analizar cómo diferencia un ítem de Psicometría entre los sujetos que han resultado aptos y no-aptos. También podemos utilizar otros criterios externos como el género, o características socio-demográficas.

Así por ejemplo, imaginemos que deseamos conocer si el ítem 5 del último examen de Psicometría discrimina adecuadamente entre los aptos y los no-aptos. En primer lugar, habrá que ordenar los datos en una tabla de contingencia 2 × 2 tal y como se muestra a continuación, donde 1 indica que se acierta el ítem o se supera el criterio, y 0 que se falla el ítem o que no se supera el criterio.

TABLA 8.6
Tabla para el cálculo de Φ

		Ítem (X)		
		1	0	
Criterio (Y)	1	a	b	(a + b)
	0	c	d	(c + d)
		(a + c)	(b + d)	N

En la tabla anterior, la celdilla *a* hace referencia al número de sujetos que han acertado el ítem y que además han aprobado el examen de Psicometría. El marginal *a* + *b* es el número de sujetos que han aprobado Psicometría; mientras que el *c* + *d* son los que no lo han superado. Por otro lado, el marginal *a* + *c* son los sujetos que han acertado el ítem; y el *b* + *d* los que lo han fallado. Si dividimos los datos anteriores entre el número total de sujetos *N* obtendremos sus respectivas proporciones:

TABLA 8.7
Proporciones para el cálculo de Φ

		Ítem (X)		
		1	0	
Criterio (Y)	1	a/N = p_{xy}	b	(a + b)/N = p_y
	0	c	d	(c + d)/N = q_y
		(a + c)/N = p_x	(b + d)/N = q_x	N

Finalmente, aplicamos la siguiente ecuación, cuya formulación algebraica es homóloga a la del coeficiente de correlación de Pearson.

$$\Phi = \frac{p_{xy} - p_x p_y}{\sqrt{p_x q_x p_y q_y}} \qquad [8.7]$$

EJEMPLO:

Tras ordenar los resultados de 50 sujetos presentados al último examen de Psicometría obtenemos la tabla 8.8.

TABLA 8.8
Datos representativos para el análisis

		Ítem 5 (X)		
		1	0	
Criterio (Y)	Apto	p_{xy} 30/50 = 0,6	5	p_y 35/50 = 0,7
	No-Apto	5	10	q_y 15/50 = 0,3
		p_x 35/50 = 0,7	q_x 15/50 = 0,3	N = 50

$$\Phi = \frac{p_{xy} - p_x p_y}{\sqrt{p_x q_x p_y q_y}} = \frac{0,6 - 0,7 \times 0,7}{\sqrt{0,7 \times 0,3 \times 0,7 \times 0,3}} = 0,52$$

Se puede concluir que existe una correlación alta entre el ítem y el criterio, es decir, aquellos sujetos que aciertan el ítem suelen aprobar el examen de Psicometría.

4.2.2. Correlación biserial-puntual

Cuando el ítem es una variable dicotómica y la puntuación en el test es continua, el índice de correlación más apropiado es el biserial-puntual. Su expresión es:

$$r_{bp} = \frac{\overline{X}_A - \overline{X}_T}{S_x}\sqrt{\frac{p}{q}} \qquad [8.8]$$

donde:

$\overline{X}_A$ = media en el test de los sujetos que aciertan el ítem.

$\overline{X}_T$ = media del test.

S_x = desviación típica del test.

p = proporción de sujetos que aciertan el ítem.

q =proporción de sujetos que fallan el ítem.

Como se ha comentado anteriormente, para calcular la correlación habría que eliminar de las puntuaciones del test las del ítem en cuestión, en caso contrario se estaría incrementando artificialmente la correlación biserial-puntual. Esta corrección es aún más necesaria cuando el número de ítems es menor de 25.

EJEMPLO:

En la siguiente tabla se muestran las respuestas de 5 sujetos a 4 ítems. Calcular la correlación biserial-puntual del segundo ítem.

TABLA 8.9
Datos del ejemplo

	Ítems				*Total*	
Sujetos	1	2	3	4	*X*	*(X–i)*
A	0	1	0	1	2	1
B	1	1	0	1	3	2
C	1	1	1	1	4	3
D	0	0	0	1	1	1
E	1	1	1	0	3	2

Los sujetos que han acertado el ítem son el A, B, C y E, luego su media es:

$$\overline{X}_A = \frac{1+2+3+2}{4} = 2$$

La media total del test es:

$$\overline{X}_T = \frac{1+2+3+1+2}{5} = 1,8$$

La desviación típica de las puntuaciones del test:

$$S_x^2 = \frac{1^2+2^2+3^2+1^2+2^2}{5} - (1,8)^2 = 0,56$$

$$S_x = \sqrt{0,56} = 0,75$$

La proporción de sujetos que han acertado el ítem 2 es 4/5 = 0,8; mientras la de sujetos que lo han fallado es 1/5 = 0,2.

Finalmente, la correlación biserial-puntual entre el ítem y las puntuaciones del test, descontando las del ítem es:

$$r_{bp} = \frac{\overline{X}_A - \overline{X}_T}{S_x}\sqrt{\frac{p}{q}} = \frac{2-1,8}{0,75}\sqrt{\frac{0,8}{0,2}} = 0,54$$

4.2.3. Correlación biserial

La correlación biserial está muy próxima a la biserial-puntual, pero con una diferencia importante en sus asunciones. Mientras que la anterior se aplica cuando una de las variables es intrínsicamente dicotómica y la otra continua, en la biserial se entiende que ambas variables son inherentemente continuas, aunque una de ellas se haya dicotomizado (ítem). Su expresión es:

$$r_b = \frac{\overline{X}_A - \overline{X}_T}{S_x}\frac{p}{y} \qquad [8.9]$$

Todos los símbolos se interpretan como en el caso de la correlación biserial-puntual. La única novedad viene dada por *y*, que hace referencia a la altura en la curva normal correspondiente a la puntuación típica que deja por debajo un valor de probabilidad igual a *p*. Los valores de *y* se pueden consultar en la tabla 7 del final del libro.

La correlación biserial del ítem 3, vendría dada por:

$$r_b = \frac{\overline{X}_A - \overline{X}_T}{S_x}\frac{p}{y} = \frac{2,5-2,2}{0,75}\left(\frac{0,4}{0,3863}\right) = 0,41$$

Para obtener la *y,* dado que el valor $p = 0,40$ no aparece en la primera columna de la tabla 7, hemos buscado el valor de 0,60 (su complementario, es decir, q) que lleva asociada una $y = 0,3863$ (columna F). Una vez conocido el valor de *y,* que ha de ser el mismo para p y q, basta calcular 0,40/0,3863 para obtener el valor del quebrado (p/y) que coincide con el que aparece en la columna E (última columna) de la tabla 7 para una $p = 0,40$ (antepenúltima columna).

Hay que destacar que la r_b es una estimación de la correlación de Pearson, y por tanto es posible hallar valores mayores que 1, especialmente, cuando alguna de las variables no es normal.

La relación entre r_{bp} y r_b viene dada por:

$$r_{bp} = r_b \frac{y}{\sqrt{pq}} \qquad [8.10]$$

Dado que el valor de *y* es siempre menos que $\sqrt{pq}$ el valor de la correlación biserial será mayor que el de la biserial-puntual. Esta diferencia será moderada en ítems de dificultad media, y se incrementará en ítems de dificultad alta y baja (Martínez-Arias, Hernández y Hernández, 2006). El lector interesado, puede comprobar la equivalencia entre r_{bp} y r_b a partir de la ecuación 8.10.

4.3. Discriminación en los ítems de actitudes

Si retomamos lo visto en el tema 3, los ítems de actitudes se caracterizan porque no existen respuestas correctas o incorrectas, sino que el sujeto ha de situarse en el continuo establecido en función del grado del atributo medido. Teniendo esto presente, y habiendo considerado que la discriminación se había definido como la correlación entre las puntuaciones del ítem y las del test, es fácil deducir que un procedimiento para estimar la discriminación de los ítems de actitudes pasa por calcular la correlación entre ambos. En este caso, al tratarse de ítems que no son dicotómicos el coeficiente de correlación adecuado sería el de Pearson. Este coeficiente de correlación, también se puede interpretar como un *Índice de Homogeneidad (IH).* Indica hasta qué punto el ítem está midiendo la misma dimensión, o en este caso actitud, que el resto de los ítems de la escala. Como norma general, aquellos ítems cuyo *IH* esté por debajo de 0,20 se han de eliminar de la escala resultante (Barbero, 2007).

$$R_{jx} = \frac{N\sum JX - \sum J\sum X}{\sqrt{\left[N\sum J^2 - \left(\sum J\right)^2\right]\left[N\sum X^2 - \left(\sum X\right)^2\right]}} = \frac{COV(JX)}{S_j S_x} \qquad [8.11]$$

donde :

N = número de sujetos de la muestra.

$\sum J$ = suma de las puntuaciones de los sujetos en el elemento *j*.

$\sum X$ = suma de las puntuaciones de los sujetos en la escala total.

R_{jx} = correlación entre las puntuaciones obtenidas por los sujetos en el elemento *j* y en la escala total.

Al igual que en los casos anteriores, es necesario tener en cuenta que si las puntuaciones del ítem están contando a la hora de calcular la puntuación total del test, habría que aplicar una corrección. Como ya vimos, dicha corrección puede implicar, simplemente, descontar de la puntuación total la del ítem o aplicar la siguiente fórmula:

$$R_{j(x-j)} = \frac{R_{jx}S_x - S_j}{\sqrt{S_x^2 + S_j^2 - 2R_{jx}S_xS_j}} \qquad [8.12]$$

Basándonos en el propio concepto de discriminación, otro procedimiento extremadamente útil (aunque menos eficiente que el anterior porque no utiliza toda la muestra) para averiguar si un ítem diferencia entre grupos extremos de actitud consiste en calcular si la media en el ítem de los sujetos con puntuaciones más altas en el test total es estadísticamente superior a la media de los sujetos con puntuaciones más bajas. Para establecer los grupos altos y bajos de actitud se suele utilizar al 25% (o 27%) de los sujetos con mejores puntuaciones y al 25% (o 27%) con puntuaciones más bajas. Una vez establecidos los grupos se procede a calcular si su diferencia de medias es estadísticamente significativa mediante la prueba de *T de Student* (Barbero, 2007):

$$T = \frac{\overline{X}_{sj} - \overline{X}_{ij}}{\sqrt{\frac{(n_s-1)S_{sj}^2 + (n_i-1)S_{ij}^2}{n_s+n_i-2}\left[\frac{1}{n_s}+\frac{1}{n_i}\right]}} \qquad [8.13]$$

donde:

$\overline{X}_{sj}$ = media de las puntuaciones obtenidas en el ítem por el 25% de los sujetos que obtuvieron puntuaciones más altas en el test.

$\overline{X}_{ij}$ = media de las puntuaciones obtenidas en el ítem por el 25% de los sujetos que obtuvieron puntuaciones más bajas en el test.

S^2_{sj} = varianza de las puntuaciones obtenidas en el ítem por el 25% de los sujetos que obtuvieron puntuaciones más altas en el test.

S^2_{ij} = varianza de las puntuaciones obtenidas en el ítem por el 25% de los sujetos que obtuvieron puntuaciones más bajas en el test.

n_s y n_i = número de sujetos con conforman respectivamente el grupo superior e inferior.

La *T de Student* obtenida se distribuye con $(n_s + n_i - 2)$ grados de libertad. La hipótesis nula que se pone a prueba es que las medias de ambos grupos son iguales. En tanto que, para un determinado nivel de confianza, obtengamos un valor empírico de *T* superior al teórico (se consulta en la tabla correspondiente) tendríamos que rechazar la H_o a favor de la hipótesis alternativa que establece que la media del grupo superior es mayor que la del inferior (contraste unilateral).

EJEMPLO:

Las respuestas de 5 sujetos a 4 ítems de actitudes se muestran en la tabla 8.10. Calcular la discriminación del elemento número cuatro (X_4) mediante la correlación de Pearson. Y la del elemento número 2 mediante la prueba *T de Student*.

TABLA 8.10
Datos del ejemplo

	Ítems							
Sujetos	X_1	X_2	X_3	X_4	Total X_T	X_4X_T	X_4^2	X_T^2
A	2	4	4	3	13	39	9	169
B	3	4	3	5	15	75	25	225
C	5	2	4	3	14	42	9	196
D	3	5	2	4	14	56	16	196
E	4	5	2	5	16	80	25	256
				20	72	292	84	1042

La correlación, o *IH* entre el elemento 4 y la puntuación total del test será:

$$R_{jx} = \frac{N\sum JX - \sum J\sum X}{\sqrt{\left[N\sum J^2 - \left(\sum J\right)^2\right]\left[N\sum X^2 - \left(\sum X\right)^2\right]}} = \frac{5\cdot 292 - 20\cdot 72}{\sqrt{\left[5\cdot 84 - 20^2\right]\left[5\cdot 1042 - 72^2\right]}} = 0{,}88$$

El inconveniente es que el resultado así obtenido está artificialmente inflado dado que en X_T está incluida la puntuación de X_4. Así que es necesario aplicar la fórmula de corrección.

La varianzas y desviaciones típicas de X_4 y X_T son:

$$S^2_{x_4} = \frac{3^2 + 5^2 + 3^2 + 4^2 + 5^2}{5} - (4)^2 = 0{,}80$$

$$S_{x_4} = \sqrt{0{,}80} = 0{,}89$$

$$S^2_{x_T} = \frac{13^2 + 15^2 + 14^2 + 14^2 + 16^2}{5} - (14{,}4)^2 = 1{,}04$$

$$S_{x_T} = \sqrt{1{,}04} = 1{,}02$$

$$R_{j(x-j)} = \frac{R_{jx}S_x - S_j}{\sqrt{S_x^2 + S_j^2 - 2R_{jx}S_xS_j}} = \frac{0{,}88\cdot 1{,}02 - 0{,}89}{\sqrt{1{,}04 + 0{,}80 - 2\cdot 0{,}88\cdot 1{,}02\cdot 0{,}89}} = 0{,}01$$

No debe sorprender que cuando se utiliza la fórmula de corrección, de 0,88 (un muy buen *IH*) hemos pasado a obtener un *IH* próximo a cero. Ello se debe a que el número de elementos que hemos empleado en el ejemplo es muy pequeño. A medida que el número de ítems aumenta, el efecto expuesto disminuye porque la influencia de las puntuaciones del ítem en la puntuación total es cada vez menor. De tal forma que cuando estemos trabajando con más de 25 ítems los resultados serán muy próximos. Obsérvese por tanto, la importancia de sustraer la puntuación del ítem de la puntuación total del test cuando calculamos su correlación. Este ejemplo, es absolutamente generalizable a los ítems de aptitudes.

Si la escala tuviera un número de ítems adecuados y hubiéramos obtenido estos mismos resultados en el ítem 4, la conclusión sobre su idoneidad indicaría la necesidad de eliminarlo de la escala dado que su *IH* corregido es próximo a cero. El ítem 4 no contribuye a medir el mismo rasgo que la escala total.

Para calcular la discriminación del elemento número 2 mediante *T de Student*, tendríamos que utilizar al 25% de los sujetos que han obtenido puntuaciones más altas para conformar el grupo superior; y el 25% de los que han presentado puntuaciones más bajas para el grupo inferior. Teniendo en cuenta que sólo se trata de un ejemplo, por motivos didácticos y para ilustrar el procedimiento vamos a utilizar a los dos sujetos con puntuaciones más altas y más bajas en X_T.

En nuestro ejemplo, los dos sujetos que han puntuado más alto en la escala han sido el B (15) y el E (16); mientras que los que han obtenido puntuaciones más bajas son el A (13) y el C (14). Las puntuaciones de dichos sujetos en el ítem 2, así como las medias y varianzas para ambos grupos son:

	Sujeto	X_2
Grupo superior	E	5
	B	4

$\overline{X}_s = 4,5;\ S_s^2 = 0,25$

	Sujeto	X_T
Grupo inferior	A	4
	C	2

$\overline{X}_I = 3;\ S_i^2 = 1$

Aplicamos la prueba de *T de Student*

$$T = \frac{\overline{X}_{sj} - \overline{X}_{ij}}{\sqrt{\frac{(n_s - 1)S_{sj}^2 + (n_i - 1)S_{ij}^2}{n_s + n_i - 2}\left[\frac{1}{n_s} + \frac{1}{n_i}\right]}} = \frac{4,5 - 3}{\sqrt{\frac{(2-1)0,25 + (2-1)1}{2+2-2}\left[\frac{1}{2} + \frac{1}{2}\right]}} = 1,9$$

El valor empírico obtenido es de 1,9. Para un NC del 95% el valor teórico que encontramos en las tablas para 2 grados de libertad (2 + 2 – 2) es de 2,92. Dado que el valor empírico obtenido en los datos de nuestra muestra es menor que el teórico, deberíamos aceptar la hipótesis nula que establece que la medida para el grupo superior no es significativamente mayor, es decir, el ítem no discrimina adecuadamente.

Estos resultados hay que interpretarlos bajo la óptica de que se trata de un ejemplo didáctico ya que para poder aplicar la *T de Student* las puntuaciones del ítem y las de la escala total han de distribuirse normalmente y sus varianzas iguales. Si no fuera este el caso, en lugar de la *T de Student* habría que aplicar alguna prueba no paramétrica para calcular la diferencia de medias (*U de Mann-Whitney*, por ejemplo).

4.4. Factores que afectan a la discriminación

4.4.1. Variabilidad

En páginas anteriores habíamos resaltado lo importante que es la presencia de variabilidad en las respuestas de los sujetos a los ítems, es decir, que sean acertados y fallados por sujetos con distinto nivel en la variable medida. Cuando la varianza de un ítem es cero, implica que todos los sujetos han respondido lo mismo, si te tratara de un ítem de un test de aptitudes todos los sujetos lo habrían acertado o fallado; cuando se trata de un ítem de un test de actitudes, personalidad, etc., donde no hay respuestas correcta o incorrectas, un ítem con varianza cero implicaría que todos los sujetos han elegido la misma alternativa de respuesta. Y cuando esto ocurre el ítem no presenta ningún poder discriminativo, dado que si su varianza es igual a cero, entonces su correlación con las puntuaciones del test también es cero (véase figura 8.3 y ecuación 8.11).

La relación entre la variabilidad del test y la discriminación de los ítems se puede formular algebraicamente:

$$S_x = \sum_{j=1}^{n} S_j r_{jx} \qquad [8.14]$$

donde:

S_x = desviación típica del test.

S_j = desviación típica del ítem.

r_{jx} = índice de discriminación del ítem j.

Es decir, la desviación típica del test puede descomponerse en el sumatorio del producto de las desviaciones típicas de los ítems por sus correlaciones con el test.

Si el test estuviera compuesto por ítems dicotómicos, dado que la varianza de una variable dicotómica es igual a la proporción de aciertos por la proporción de fallos, la ecuación 8.14 deriva en:

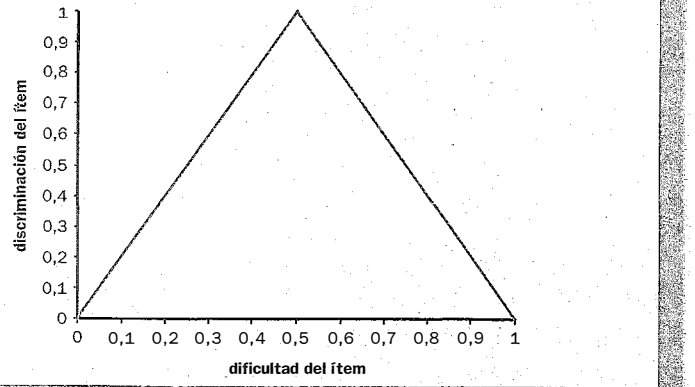
$$S_x^2 = \sum_{j=1}^{n} p_j q_j r_{jx}^2 \Rightarrow S_x = \sqrt{\sum_{j=1}^{n} p_j q_j r_{jx}^2} \qquad [8.15]$$

En la ecuación 8.15, es donde mejor se puede apreciar que para maximizar la capacidad discriminativa de un test habrá que considerar conjuntamente tanto la dificultad (p_j) como la discriminación (r_{jx}) de sus ítems. Esto se consigue cuando la discriminación sea máxima ($r_{jx} = 1$) y su dificultad media ($p = 0,5$) (comprobar en la ecuación 8.15).

4.4.2. Dificultad del ítem

Un ítem alcanza su máximo poder discriminativo cuando su dificultad es media. Implícitamente, esta idea ya se ha expuesto cuando se relacionaba la dificultad con la varianza del test. Concretamente, se dijo que la varianza sería máxima cuando su dificultad fuera media (p = 0,5 en ítems dicotómicos), y justamente en el epígrafe anterior se ha comentado que la varianza del ítem es clave para que éste discrimine. Luego para optimizar la discriminación habrá que tener muy en cuenta la dificultad del ítem. En la figura 8.5 se relacionan los valores de dificultad y discriminación.

FIGURA 8.5
Relación entre dificultad y discriminación



4.4.3. Dimensionalidad del test

La dimensionalidad de un test hace referencia al número de conceptos o constructos que se están midiendo. Su estudio está estrechamente relacionado con la validez de constructo y para su examen la técnica más utilizada es el Análisis Factorial, del que ya se apuntó algo en el tema 6 sobre Validez.

Cuando se construye un test, se trata de que sólo mida un único concepto, es decir, que sea unidimensional. Si tras someter el test a un Análisis Factorial encontráramos varias dimensiones subyacentes, implicaría la existencia de distintas escalas, lo que sería similar a una batería de test que

mide tantos aspectos como escalas o dimensiones distintas hubiera. Si fuera este el caso, la correlación entre las puntuaciones en el test y las del ítem se verá afectada a la baja, y tanto más cuanto más dimensiones contenga el test.

En tests multidimensionales, la discriminación de los ítems hay que estimarla única y exclusivamente considerando el conjunto de ítems que se asocian a cada dimensión o concepto. Si no es así, podemos llegar a desechar ítems que en su dimensión presenten gran poder discriminativo.

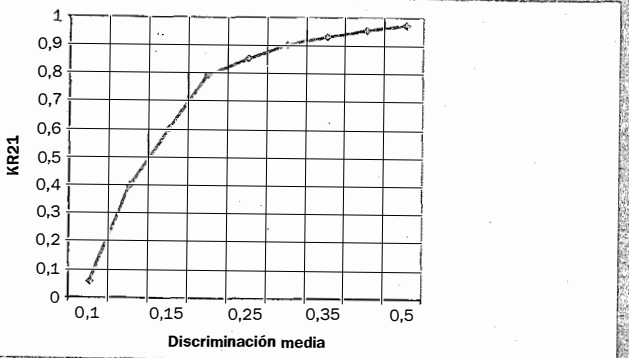
4.4.4. Fiabilidad del test

Si la discriminación se define como la correlación entre las puntuaciones obtenidas por los sujetos en el ítem y las del test, entonces fiabilidad y discriminación han de estar íntimamente relacionados. Tan es así que es posible expresar el coeficiente *alpha de Cronbach* a partir de la discriminación de los ítems (también de su dificultad). Para ello, basta con sustituir S_x por $\Sigma S_j r_{jx}$ (véase ecuación 8.14).

$$\alpha = \frac{n}{n-1}\left(1-\frac{\sum_{j=1}^{n} S_j^2}{S_x^2}\right) = \frac{n}{n-1}\left(1-\frac{\sum_{j=1}^{n} S_j^2}{\left[\sum_{j=1}^{n} S_j r_{jx}\right]^2}\right) \qquad [8.16]$$

Valores pequeños en la discriminación de los ítems suelen estar asociados con tests poco fiables (comprobar en la ecuación 8.16). Esta relación queda representada en la figura 8.6, que relaciona el coeficiente KR_{21} con la discriminación media de un test compuesto por 100 ítems cuya p = 0,8 en todos ellos. A medida que aumenta la discriminación media del test, el coeficiente de fiabilidad se incrementa, de esta forma para un valor de 0,15 de discriminación media KR_{21} es 0,60, mientras que para un valor de 0,29 la fiabilidad alcanza un coeficiente de 0,90.

FIGURA 8.6
Relación entre fiabilidad y discriminación



Finalmente, habría que destacar que aunque técnicamente sea factible obtener muy buenos ítems desde un punto de vista psicométrico mediante la combinación óptima de los factores anteriores, el examen definitivo para un ítem implica que los sujetos más competentes elijan la alternativa correcta en mayor proporción que los sujetos menos competentes en el dominio de interés.

5. ÍNDICES DE FIABILIDAD Y VALIDEZ DE LOS ÍTEMS

5.1. Índice de fiabilidad

Se utiliza para cuantificar el grado en que el ítem en cuestión está midiendo con precisión el atributo de interés. Su formulación matemática la podemos encontrar en la fórmula 8.17, concretamente:

$$IF = S_j ID_j \qquad [8.17]$$

donde:

S_j = desviación típica de las puntuaciones en el ítem.

ID_j = índice de discriminación del ítem.

Cuando se utiliza algún coeficiente de correlación para calcular la discriminación de los ítems entonces:

$$IF = S_j r_{jx} \qquad [8.18]$$

que justamente es uno de los denominadores de la ecuación 8.16. Por tanto, el sumatorio al cuadrado de los *IF* de los ítems coincide con la varianza de las puntuaciones de los sujetos en el test.

Observando la ecuación 8.16, es fácil entender la relación directa entre la fiabilidad de los ítems y la del test. En la medida que seleccionemos los ítems con mayor *IF*, mayor será su sumatorio ($\Sigma S_j r_{jx} = \Sigma IF$), y por ende mejor la fiabilidad del test.

5.2. Índice de validez

Tal y como se ha visto en el epígrafe titulado «*la validación referida al criterio*» del tema 6, la validez implica correlacionar las puntuaciones del test con algún criterio externo de interés. Análogamente en el caso de un ítem concreto, implicará correlacionar las puntuaciones obtenidas por una muestra de sujetos en el ítem con las puntuaciones obtenidas por los mismos sujetos en algún criterio externo de interés. Esto sirve para determinar hasta qué punto cada uno de los ítems de un test contribuye a realizar con éxito predicciones sobre dicho criterio externo.

$$IV = S_j r_{jy} \qquad [8.19]$$

En el caso de que el criterio sea una variable continua y el ítem una variable dicotómica, la correlación a utilizar sería la biserial puntual; pero ahora no es necesario descontar de la puntuación total del criterio externo la del ítem ya que ésta no está incluida y, por lo tanto, no contribuye de ninguna manera en su cómputo.

$$IV = S_j r_{bp_{jy}} \qquad [8.20]$$

Si anteriormente habíamos expuesto que la fiabilidad del test depende de los *IF* de los ítems, la validez del test también puede expresarse en función de los *IV* de los ítems, de manera que cuanto mayores sean los *IV* de los ítems, más optimizarán la validez del test (Muñiz, 2003).

$$r_{xy} = \frac{\sum_{j=1}^{n} S_j r_{jy}}{\sum_{j=1}^{n} S_j r_{jx}} = \frac{\sum_{j=1}^{n} IV_j}{\sum_{j=1}^{n} IF_j} \qquad [8.21]$$

La ecuación 8.21 es muy importante porque permite ver cómo la validez del test se puede estimar a partir de la discriminación de cada uno de los ítems (r_{jx}), de su validez (r_{jy}), y de su variabilidad $\left(S_j = \sqrt{p_j q_j}\right)$.

Considerando conjuntamente las ecuaciones 8.16 y 8.21, encontramos una paradoja en la selección de los ítems. Es decir, si queremos seleccionar ítems para maximizar la fiabilidad del test tendremos que elegir aquellos cuyo índice de discriminación (r_{jx}) sea alto (ecuación 8.16); pero esta política nos llevaría a reducir el coeficiente de validez del test (ecuación 8.21) porque ésta aumenta a medida que los índices de validez son elevados y los de fiabilidad bajos. Por tanto, si deseamos incrementar la validez o la fiabilidad del test a partir de la selección de los ítems, se plantea una difícil cuestión que ha de ser sometida al criterio del investigador o del constructor del test.

6. ANÁLISIS DE DISTRACTORES

Si el análisis de la alternativa correcta (todo lo anterior se basa en ello) es importante para la mejora de la calidad de los ítems, igualmente relevante resulta el análisis de los distractores o respuestas incorrectas. Este análisis implica indagar en la distribución de los sujetos a lo largo de los distractores, lo que permite, entre otras cosas, detectar posibles motivos de la baja discriminación de algún ítem, o constatar que algunas alternativas no son seleccionadas por nadie.

Para llevar a cabo este tipo de análisis y comprobar que los distractores utilizados son correctos hay que seguir los siguientes pasos:

1. Controlar que todas las opciones incorrectas sean elegidas por un mínimo de sujetos.
2. A ser posible, que sean equiprobables, es decir, que sean igualmente atractivas para las personas que no conocen la respuesta correcta.
3. Que el rendimiento en el test de los sujetos que han seleccionado cada alternativa incorrecta sea menor al de los sujetos que han seleccionado la correcta.
4. En relación con el punto anterior, es de esperar que a medida que aumente el nivel de aptitud de los sujetos, el porcentaje de ellos que seleccionen las alternativas incorrectas disminuya, y viceversa, a medida que disminuya el nivel de aptitud de los sujetos el porcentaje de los que seleccionen los distractores aumente.

6.1. Equiprobabilidad de los distractores

Una forma de comprobar la equiprobabilidad de los distractores es mediante la aplicación de la prueba de independencia de χ^2 (García-Cueto, 2005).

$$\chi^2 = \sum_{j=1}^{k} \frac{(FT - FO)^2}{FT} \qquad [8.22]$$

donde:

FT = frecuencias teóricas.

FO = frecuencias observadas.

Los grados de libertad son (*k*–*1*), donde *k* es el número de alternativas incorrectas. La hipótesis nula a poner a prueba es que *FT* = *FO*, que significa que para los sujetos que no conocen la respuesta correcta la elección de cualquiera de los distractores es igualmente atractiva.

Retomando los datos del ejemplo de la tabla 8.4, si queremos determinar si las alternativas incorrectas son igualmente atractivas, tendremos que aplicar la ecuación 8.22.

TABLA 8.11
Distribución de distractores

	A	B*	C
27% superior	19	53	28
46% intermedio	52	70	48
27% inferior	65	19	16
TOTAL	136	—	92

En nuestro ejemplo la *FT* será igual a (136 + 92)/2 = 114. Cada distractor ha de ser seleccionado por 114 sujetos, que en este ejemplo equivale a la mitad de los que han respondido incorrectamente al ítem. La *FO* es la que aparece en la última fila de la tabla (nótese que la alternativa *B* no

la consideramos puesto que es la alternativa correcta y estamos analizando la equiprobabilidad de los distractores).

$$\chi^2 = \sum_{j=1}^{n} \frac{(FT - FO)^2}{FT} = \frac{(114-136)^2 + (114-92)^2}{114} = \frac{968}{114} = 8,49$$

Si acudimos a las tablas de χ^2, encontramos que para 1 grado de libertad y un N.C del 95% el valor teórico de χ^2 es 3,84. Dado que el valor empírico obtenido (8,49) es mayor que el teórico (3,84) la conclusión es que las alternativas incorrectas no son igualmente atractivas para todos los sujetos, aunque sean elegidas por un mínimo del 10%.

6.2. Poder discriminativo de los distractores

Los puntos dos y tres anteriores están directamente relacionados con el concepto de discriminación. Si son buenos distractores, lo lógico es que discriminen en sentido contrario a como lo hace la opción correcta. Es decir, si se espera que la correlación entre las puntuaciones del test y la opción correcta sea positiva, y cuanto más mejor, lo esperable de un buen distractor es que su correlación sea negativa. Lo que implica que a medida que aumenta el nivel de aptitud de los sujetos la proporción de sujetos que elige el distractor disminuya.

Para ilustrar gráficamente lo comentado anteriormente, en las figuras 8.7 y 8.8 se presentan ejemplos reales de dos ítems. En el caso de la figura 8.7 se observa que a medida que aumenta la nota de los sujetos (desde no-apto a notable) la opción correcta (*a*) es seleccionada cada vez en mayor proporción, lo que redunda en una correlación positiva entre la opción correcta y las notas de los sujetos en el test (*discriminación positiva*). En los distractores (*b y c*), la tendencia es la contraria. En niveles de aptitud bajo, son igualmente seleccionadas, y a medida que el nivel de aptitud aumenta la eligen cada vez menos sujetos (*discriminación negativa*). En resumen, las opciones incorrectas discriminan en sentido contrario que la correcta.

En la figura 8.8, se presenta el caso de un mal ítem. Es malo porque la opción correcta *(a)* es selecciona aproximadamente en la misma proporción por sujetos poco competentes y muy competentes (discriminación baja o próxima a cero). Lo mismo ocurre con las alternativas incorrectas, que son seleccionadas indistintamente por no-aptos, aprobados y notables (discriminación próxima a cero), además el distractor *C* apenas es elegido por nadie, lo que significa que es fácilmente identificado como incorrecto por cualquier sujeto y por tanto tendría que revisarse.

FIGURA 8.7
Ejemplo de un buen ítem

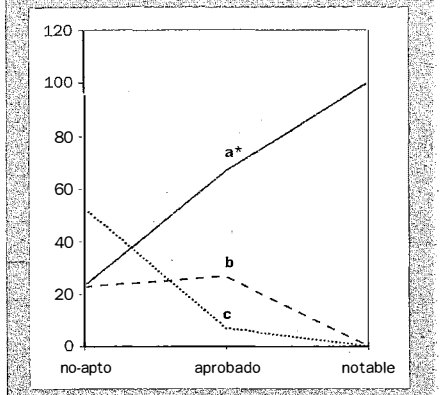
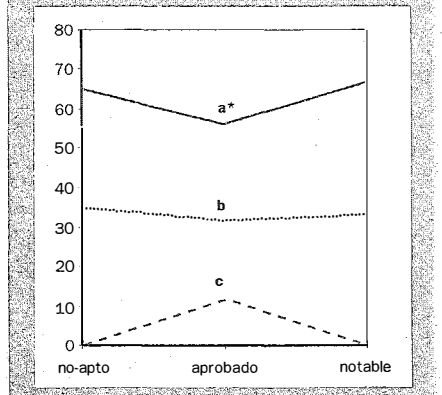


FIGURA 8.8
Ejemplo de un mal ítem



Para cuantificar el poder discriminativo de las alternativas incorrectas, nos valemos de la correlación. Dependiendo del tipo de variable utilizaremos la biserial, biserial-puntual, *phi* o Pearson.

EJEMPLO:

En la tabla 8.12 se muestran las respuestas de 5 sujetos a 4 ítems. Entre parétesis se muestra la alternativa seleccionada por cada sujeto y la alternativa correcta con asterisco. Calcular la discriminación del distractor *b* del ítem 3.

TABLA 8.12
Datos del ejemplo

	Ítems				Total	
Sujetos	1 (a*)	2 (b*)	3 (a*)	4 (c*)	X	(X–i)
A	0 (b)	1	0 (b)	1	2	2
B	1	1	0 (b)	1	3	3
C	1	1	1	1	4	3
D	0 (c)	0 (a)	0 (b)	1	1	1
E	1	1	1	0 (b)	3	2

Los sujetos que han seleccionado la alternativa *b*, que es incorrecta, en el ítem 3 han sido el A, B y D, luego la media de estos sujetos en el test después de eliminar la puntuación correspondiente al ítem analizado, es:

$$\bar{X}_A = \frac{2+3+1}{3} = 2$$

La media total del test descontando de las puntuaciones obtenidas por los sujetos, la correspondiente al ítem 3 es:

$$\bar{X}_{T-i} = \frac{2+3+3+1+2}{5} = 2,2$$

La desviación típica de las puntuaciones correspondientes a *(X–i)*

$$S^2_{x-i} = \frac{2^2+3^2+3^2+1^2+2^2}{5} - (2,2)^2 = 0,56$$
$$S_{x-i} = \sqrt{0,56} = 0,75$$

La proporción de sujetos que han acertado el ítem 3 es 2/5 = 0,40; mientras la de los sujetos que lo han fallado es 3/5 = 0,60.

Finalmente, la correlación biserial-puntual entre la alternativa incorrecta *b* y las puntuaciones del test, descontando las del ítem es:

$$r_{bp} = \frac{\bar{X}_A - \bar{X}_{T-i}}{S_{x-i}}\sqrt{\frac{p}{q}} = \frac{2-2,2}{0,75}\sqrt{\frac{0,40}{0,60}} = -0,22$$

Nota: Téngase en cuenta que al ser la alternativa incorrecta la puntuación de estos sujetos en el ítem es 0 y, por lo tanto no es necesario eliminar nada del test total.

El resultado obtenido es –0,22, lo que indica que este distractor discrimina en sentido contrario a como lo hace la alternativa correcta, tal y cómo cabría esperar de un buen distractor.

A veces, en el análisis de los ítems basta con una simple inspección visual de la distribución de respuestas de los sujetos a las distintas alternativas. Así por ejemplo, en la tabla 8.13 se muestra el número de sujetos de los grupos extremos de aptitud que han seleccionado cada una de las alternativas de un ítem, donde la *c* es la correcta. Para cada alternativa se muestra la proporción de sujetos que la han seleccionado (*p*), la media en el test de los sujetos que han seleccionado cada alternativa (*media*) y el índice de discriminación (r_{bp}) de todas las opciones.

TABLA 8.13
Análisis de distractores

		A	B	C*
Nivel de aptitud	*Superior*	20	25	55
	Inferior	40	35	25
Estadísticos	*p*	0,28	0,50	0,22
	Media	5	10	9
	r_{bp}	–0.20	0.18	0.29

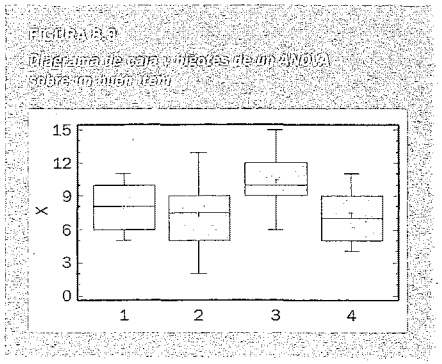
Considerando los criterios anteriores, vemos que la alternativa correcta es mayoritariamente elegida por sujetos competentes, lo que se refleja en un índice de discriminación positivo.

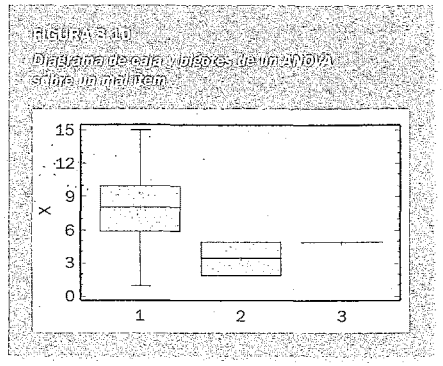
La alternativa incorrecta *A*, en principio ha sido elegida por un mínimo aceptable de sujetos (28%), y es seleccionada en mayor proporción por los sujetos menos competentes que por los competentes. Además la media en el test de los sujetos que la han seleccionado es menor que la media de los que han seleccionado la alternativa correcta lo que es coherente con el índice de discriminación negativo que presenta.

Finalmente, el distractor *B* ha de ser revisado dado que es elegido como correcto por los sujetos con mejores puntuaciones en el test. Además, ha sido la opción más seleccionada (50%), su discriminación es positiva, y la media de los sujetos que la han seleccionado es mayor que la de los sujetos que han optado por la alternativa correcta.

En el análisis de distractores aun podemos ir mucho más allá y recurrir a la inferencia estadística. En buena lógica, la media en el test de los sujetos que optan por la alternativa correcta ha de ser mayor que la media de los sujetos que han elegido cada una de las incorrectas. Este extremo se puede poner a prueba mediante un Análisis de la Varianza, en el que la variable independiente, o factor, sea cada uno de los ítems con tantos niveles como alternativas de respuesta; y la variable dependiente sea la puntuación directa de los sujetos en el test (*X* = suma de los ítems acertados correctamente). Si los distractores discriminan adecuadamente se supone que deberíamos encontrar diferencias estadísticamente significativas entre la alternativa correcta y el resto de alternativas. De la misma manera, si las alternativas incorrectas fueran equiprobables, no se deberían encontrar di-

ferencias estadísticamente significativas entre ellas. Un simple diagrama de caja y bigotes nos puede servir para ilustrarlo. A continuación a título de ejemplo, se muestra el diagrama de cajas y bigotes de un ítem cuyas 4 alternativas funcionan correctamente, y el diagrama de otro ítem (también de 4 alternativas) que tendría que ser sometido a un profundo proceso de revisión.





El diagrama de caja y bigotes del ítem representado en la figura 8.9, presenta resultados coherentes con la hipótesis de que los distractores funcionan adecuadamente. De esta forma se aprecia que la media de los sujetos que han seleccionado la opción correcta (*3*) es más alta en el test que la de los que han seleccionado el resto de las opciones. A su vez se aprecia que la dispersión de los sujetos que han seleccionado la alternativa correcta apenas se solapa con los que han seleccionado las opciones incorrectas *2* y *4*, no ocurriendo lo mismo con la opción 1, que en cierta medida podría estar confundiendo a algunos de los sujetos con buenas puntuaciones en el test. En este mismo sentido también se observa que los tres distractores atraen aproximadamente de la misma forma a los sujetos con una aptitud media o baja, por lo tanto podemos concluir que están funcionando correctamente.

En el diagrama de cajas y bigotes de la figura 8.10 se observa una gran inconsistencia en las respuestas de los sujetos a las distintas opciones de respuesta. En este caso la opción *4* no ha sido seleccionada por nadie lo cual indica que es claramente identificada como falsa (no aparece en el diagrama), y por tanto tendría que revisarse ya que no atrae a los sujetos que en principio no tienen por qué responder correctamente al ítem. Además la opción incorrecta 3, sólo ha sido seleccionada por un sujeto cuya puntuación ha sido baja en el test, por tanto no funciona correctamente como distractor ya que no atrae a sujetos con un nivel bajo o medio. La opción incorrecta 2 presenta una variabilidad muy pequeña. La opción correcta (1), ha sido respondida indistintamente por sujetos de baja y alta aptitud lo que la invalida para diferenciar a sujetos con distintos niveles en la variable medida (los bigotes ocupan prácticamente todo el rango de *X*). Según estos resultados se puede decir que las opciones de este ítem tendrían que ser revisadas ya que no sirven para diferenciar a sujetos con distinto nivel de aptitud.

7. FUNCIONAMIENTO DIFERENCIAL DE LOS ÍTEMS (FDI)

Otro aspecto a evaluar dentro del análisis de ítems, es si de manera sistemática sujetos de distintos grupos de pertenencia pero con el mismo nivel en el rasgo medido tienen distintas probabilidades de éxito en el ítem en cuestión (Shultz y Whitney, 2005). A esta circunstancia se la conoce como *funcionamiento diferencial de los ítems* (FDI), reservando la palabra *sesgo* para la interpretación de las causas de dicho funcionamiento diferencial. Por el contrario, si dichas diferencias son debidas a una diferencia real en la variable medida y no a fuentes sistemáticas de variación entonces hablamos de *impacto* (Ackerman, 1992).

Conviene aclarar los tres conceptos presentados en el párrafo anterior; *sesgo, FDI, e impacto.*

En palabras de Muñiz (p. 236, 2001) «*Un metro estará sistemáticamente sesgado si no proporciona la misma medida para dos objetos o clases de objetos que de hecho miden lo mismo, sino que sistemáticamente perjudica a uno de ellos*». En nuestro contexto, un ítem estará sesgado si sujetos igualmente hábiles no tienen la misma probabilidad de acertarlo por el hecho de pertenecer a subpoblaciones distintas. El concepto de sesgo se reserva para el estudio del motivo o causa por el que el ítem beneficia a unos sujetos frente a otros con la misma aptitud. Este aspecto entronca directamente con la validez ya que implica un error sistemático (siempre en la misma dirección), y dentro de la validez, concretamente, con la de constructo porque un ítem sesgado implica que no está midiendo el mismo rasgo en ambas subpoblaciones. En este caso, el rendimiento de alguna de las subpoblaciones está afectado por alguna otra variable extraña distinta a la que se supone que mide el ítem.

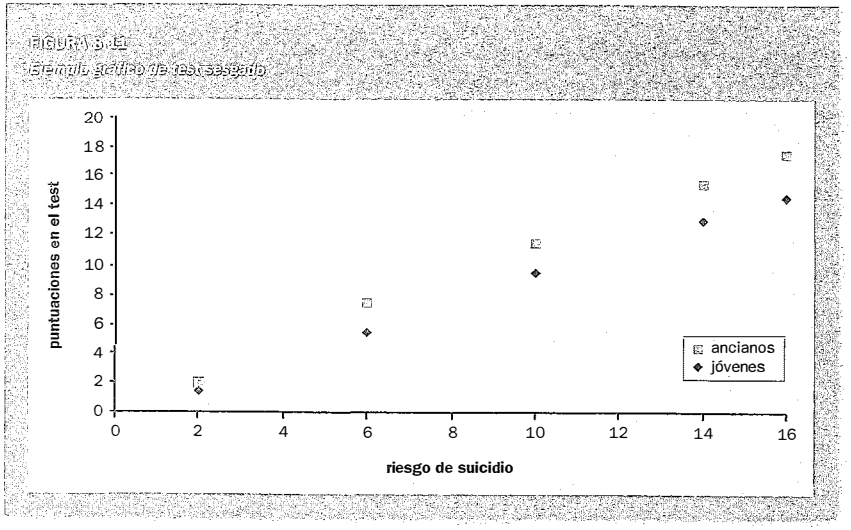
El FDI es la herramienta que utilizamos para detectar posibles ítems sesgados. Para ello, hemos de comparar el rendimiento de grupos conformados por alguna variable externa al concepto que el ítem mide (género, raza, nivel económico,...), y que sin embargo estén equiparados en cuanto a su nivel de aptitud. El FDI, simplemente detecta que un ítem funciona de manera distinta en dos grupos con el mismo nivel de aptitud (actitud, habilidad, competencia...), pero una vez detectado el fenómeno, no apunta posibles causas.

Reservamos el término *impacto*, para referirnos a diferencias reales entre grupos. Es absolutamente lícito que el rendimiento de dos grupos en un ítem sea distinto, y que ello se deba a diferencias en cuanto al nivel de competencia de las subpoblaciones. La distinción entre FDI e impacto, estriba en que mientras en el primero dichas diferencias no son reales (se deben a algún otro motivo distinto al nivel de aptitud), en el impacto, sencillamente, un grupo de sujetos es más hábil que otro (piénsese en un aula de un colegio que ha recibido mejor instrucción que otra).

Así por ejemplo, imaginemos que dos grupos distintos de un curso de formación continua sobre el manejo de procesadores de texto han tenido profesores distintos. El profesor del grupo A ha centrado su docencia sobre un procesador de texto denominado «*palabra*», mientras que el otro profesor (grupo B) ha impartido una docencia mucho más general dedicando bastante menos horas a «*palabra*». Al finalizar el curso se ha aplicado un test de rendimiento sobre dicho procesador, y se encuentra que el promedio de rendimiento del grupo A es mayor que el del grupo B. ¿Existe impacto o FDI? Muy probablemente, dado que el grupo A ha recibido una instrucción mucho mejor sobre «*palabra*» han desarrollado mucha más competencia que el grupo B, por lo que habrá diferencias reales, y por tanto impacto entre ambos grupos. Para descartar la presencia de FDI, tendríamos que comparar las probabilidades de éxito en cada ítem de los sujetos del grupo A y B que hayan obtenido la misma puntuación en la prueba de rendimiento sobre «*palabra*». Si los ítems no funcionan diferencialmente, entonces deberíamos encontrar las mismas posibilidades de éxito entre sujetos de ambos grupos igualados en aptitud.

Es fácil entender que nos encontramos ante un problema crucial en la construcción de tests ya que la presencia de sesgo puede tener importantes repercusiones sociales. Para ilustrar este extremo, simplemente imaginemos que un test para detectar el riesgo de suicidio entre pacientes clínicos está sesgado. El test funciona correctamente entre la población anciana, pero no entre los jóvenes. Como resultado de aplicar este test habría muchos jóvenes con un alto riesgo de suicidio que no habrían sido detectados y, por lo tanto, no habrían sido tratados adecuadamente. Como se

FIGURA 8.11
Ejemplo gráfico de test sesgado



ha apuntado anteriormente, para detectar el posible FDI habrá de comparar la probabilidad de riesgo de suicidio reportada por la prueba entre sujetos (ancianos y jóvenes) con la misma tendencia suicida. Si supiéramos a ciencia cierta el riesgo de suicidio de los sujetos, podríamos establecer varios niveles (normalmente entre 5 y 10) y comparar las puntuaciones del test entre jóvenes y ancianos en cada nivel. Es de esperar que si el test no está sesgado dichas puntuaciones sean iguales para ambos grupos.

En la figura 8.11, observamos claramente el peligro que supone utilizar este test. Sujetos con el mismo riesgo de suicidio puntúan en el test diferencialmente en función de su grupo de edad. Así por ejemplo, cuando el riesgo de suicidio es 16, los jóvenes obtienen en el test una puntuación mucho menor que los ancianos, lo que podría estar motivando que sujetos jóvenes que necesitan una atención psicológica urgente no la reciban. Precisamente, cuando menos riesgo de suicidio existe (2) es cuando el test ofrece puntuaciones más similares entre ambos grupos.

7.1. Mantel-Haenszel

Para detectar el FDI existen una amplia variedad de procedimientos estadísticos. Por su parsimonia y buenos resultados el método de Mantel-Haenszel (1959) es uno de los más utilizados y además se encuentra implementado en gran parte de las aplicaciones informáticas sobre FDI.

Para aplicar Mantel-Haenszel, en primer lugar habrá que identificar una variable que sea la posible causante del FDI. Una vez seleccionada, hemos de conformar dos grupos, uno de Referencia (GR), y otro Focal (GF). El GR suele coincidir con el grupo favorecido. Por el contrario, el GF suele ser el conformado por los sujetos perjudicados. Luego se establecen distintos niveles de aptitud tomando la puntuación empírica obtenida en el test y, finalmente, se cuenta el número de respuestas correctas e incorrectas por cada grupo (GR y GF) y nivel de habilidad *i*.

Todo lo anterior, se traduce en la siguiente hipótesis nula: un ítem no presentará FDI si el cociente entre los sujetos que aciertan el ítem y los que lo fallan es el mismo para los dos grupos en cada uno de los niveles de aptitud. Es decir:

$$H_o : \frac{A_i}{B_i} = \frac{C_i}{D_i} \quad \text{para todas las categorías}$$

donde:

A_i, B_i, C_i y D_i son las frecuencias absolutas de cada una de las categorías de habilidad *i* de la siguiente tabla de contingencia 8.14:

TABLA 3.14
Tabla de contingencia Mantel-Haenszel

	Correctas	Incorrectas	
GR	A_i	B_i	n_{Ri}
GF	C_i	D_i	n_{Fi}
	n_{1i}	n_{0i}	N_i

Una vez confeccionadas las tablas anteriores (una para cada nivel de aptitud *i*) aplicamos el estadístico de Mantel-Haenszel.

$$\alpha_{MH} = \frac{\sum_{i=1}^{n} \frac{A_i D_i}{N_i}}{\sum_{i=1}^{n} \frac{B_i C_i}{N_i}} \qquad [8.23]$$

Los valores obtenidos oscilan entre cero e infinito. Valores mayores que 1 indican que el ítem favorece al GR y menores al GF. Valores iguales a 1 o próximos indican que el ítem no presenta FDI.

EJEMPLO:

Existen indicios de que un ítem de las pruebas de acceso al PIR podría estar perjudicando a los graduados por la UNED. Para investigar esta posibilidad se han conformado 5 grupos de aptitud a partir de las puntuaciones del examen de ingreso al PIR. Utilizar el método de Mantel-Haenszel para comprobar si dicho ítem presenta FDI.

TABLA 3.15
Datos del ejemplo

A	NO – UNED (GR)		UNED (GF)	
Nota examen	*aciertos*	*fallos*	*aciertos*	*fallos*
0-4	2	7	0	9
5-10	15	51	8	51
11-15	25	48	21	80
16-20	67	14	50	35
21-35	43	8	37	10

Los datos de la tabla anterior se organizan de acuerdo con las siguientes tablas, una para cada nivel de aptitud.

Nivel I de habilidad ($0 \leq X \leq 4$)

	aciertos	fallos
GR	2	7
GF	0	9

18

Nivel II de habilidad ($5 \leq X \leq 10$)

	aciertos	fallos
GR	15	51
GF	8	51

125

Nivel III de habilidad ($11 \leq X \leq 15$)

	aciertos	fallos
GR	25	48
GF	21	80

174

Nivel IV de habilidad ($16 \leq X \leq 20$)

	aciertos	fallos
GR	67	14
GF	50	35

166

Nivel de habilidad (0 ≤ X ≤ 4)

	aciertos	fallos	
GR	43	8	
GF	37	10	
			98

Sintetizando los datos de las tablas anteriores, para facilitar los cálculos podemos construir la tabla 8.16.

TABLA 8.16
Numerador y denominador de α_{MH}

Niveles de aptitud	$(A_i \times D_i)/N_i$	$(B_i \times C_i)/N_i$
Nivel I	(2 × 9)/18 = **1**	(7 × 0)/18 = **0**
Nivel II	(15 × 51)/125 = **6,12**	(51 × 8)/125 = **3,26**
Nivel III	(25 × 80)/174 = **11,49**	(48 × 21)/174 = **5,79**
Nivel IV	(67 × 35)/166 = **14,13**	(14 × 50)/166 = **4,22**
Nivel V	(43 × 10)/98 = **4,39**	(8 × 37)/98 = **3,02**
Total	**37,13**	**16,29**

$$\alpha_{MH} = \frac{\sum_{i=1}^{n} \frac{A_i D_i}{N_i}}{\sum_{i=1}^{n} \frac{B_i C_i}{N_i}} = \frac{37,13}{16,29} = 2,28$$

A la vista de los resultados, podemos concluir que el ítem presenta FDI. El ítem perjudica sistemáticamente a los psicólogos graduados por la UNED. Por lo tanto habría que sustituirlo para evitar la discriminación observada.

8. RESUMEN

Llegados a este punto, una buena pregunta que podríamos plantear es ¿qué propiedades hacen que un ítem sea un buen instrumento de medida psicológico? Una respuesta inmediata es que un ítem es bueno cuando ayuda a mejorar el test que se pretende desarrollar. Tarea de la que se ocupa el *análisis de los ítems*, sin embargo, hay que enfatizar que este tipo de análisis proporciona información necesaria pero no suficiente acerca de la adecuación de los ítems como indicadores o conductas del dominio de interés. Es decir, si bien cualquier ítem puede presentar unos estadísticos excelentes respecto a su calidad psicométrica, podría tratarse de un elemento absolutamente irrelevante para medir el constructo de interés si no se han tenido en cuenta los objetivos de la medida, ni la relevancia y representatividad de los elementos seleccionados. En cualquier caso, las condiciones necesarias que debería satisfacer un ítem son:

1. La *dificultad* ha de ser apropiada para los sujetos a los que se les va a administrar. En líneas generales, en tests de ejecución máxima, los ítems no deben tener dificultades ni por debajo de 0,20, ni por encima de 0,80. Además, se recomienda que la mayoría de ellos presenten niveles medios de dificultad, es decir, entre 0,30 y 0,70. Ítems extremadamente fáciles, o difíciles no contribuyen a discriminar entre sujetos con distinto nivel en el rasgo medido. En ítems de actitudes, la dificultad es un parámetro al que no hay que prestarle tanta atención para mejorar la calidad de la prueba. Se traduce en el grado de actitud media de los sujetos ante el ítem, así que dependiendo de si es una actitud positiva (actitud ante el altruismo por ejemplo) o negativa (actitud ante la violencia) obtener un valor medio alto será bueno o malo respectivamente.
2. Los ítems deben *discriminar* claramente entre los grupos altos y bajos en aptitud y actitud. A veces encontramos ítems que discriminan en sentido negativo, esto es, sujetos con puntuaciones bajas en el test tienden a seleccionar la alternativa correcta en mayor proporción que los sujetos con puntuaciones altas. Esta situación suele estar indicando que, por alguna razón, los sujetos con una buena aptitud se ven atraídos por alguna opción incorrecta ambigua que, sin embargo, no resulta atractiva para los estudiantes con bajo nivel y que el propio redactor del ítem no han podido detectar. En tal caso, el ítem debería ser revisado o descartado. Cuanto más discrimine un ítem mucho mejor (por encima de 0,30 en los de aptitudes; y de 0,20 en los de actitudes).
3. Los *distractores* deben funcionar como tales. Cada alternativa incorrecta debe ser seleccionada por bastantes más sujetos con puntuaciones bajas en el test que por aquellos otros que presentan un buen nivel de aptitud, y además las alternativas incorrectas deben ser equiprobables.
4. Cuando sujetos que tienen el mismo nivel en el rasgo presentan distinta probabilidad de acertar un determinado ítem, es necesario llevar a cabo un análisis exhaustivo por si fuera un ítem que presentara funcionamiento diferencial y estuviera provocando una clara discriminación en una de las subpoblaciones estudiadas. En este caso el ítem debería ser revisado o eliminado.

9. EJERCICIOS DE AUTOEVALUACIÓN

1. Las respuestas de 10 sujetos a un ítem dicotómico de tres alternativas se muestran en la siguiente tabla, donde los 5 primeros sujetos son los que peores puntuaciones han obtenido en el test total, mientras que los 5 últimos los que más han puntuado. Calcular el índice de dificultad (*ID* e ID_c) del ítem en el grupo total (10 sujetos), en el grupo con peores puntuaciones (5 sujetos) y en el grupo con mejores puntuaciones (5 sujetos). Y el índice de discriminación del ítem.

Sujetos con peores puntuaciones			Sujetos con mejores puntuaciones		
Sujeto	*Respuestas al ítem*	*Puntuación total*	*Sujeto*	*Respuestas al ítem*	*Puntuación total*
A	0	8	*F*	1	27
B	1	12	*G*	0	28
C	0	5	*H*	1	30
D	1	10	*I*	1	27
E	0	7	*J*	1	25

2. La proporción del 25% de sujetos con mejores puntuaciones en un test de 3 elementos que acertaron el ítem 2 es del 70%, mientras que en el 25% de los que obtuvieron puntuaciones más bajas es del 32%. Con estos datos calcular el poder discriminativo del ítem 2 mediante el índice D. Seleccionada una muestra aleatoria de 5 sujetos, presentan los siguientes resultados en el test completo (entre paréntesis la opción correcta; y en cada celdilla la elegida por cada sujeto). A partir de esos datos calcular el poder discriminativo del ítem 2 utilizando para ello la correlación biserial-puntual y la biserial. Y calcular la discriminación del distractor *c* en el ítem 1.

	Ítems		
Sujetos	*1 (a)*	*2 (b)*	*3 (c)*
A	a	b	c
B	a	a	c
C	c	b	c
D	b	c	a
E	a	b	c

3. En la tabla siguiente se representan las puntuaciones dadas a un ítem, por el 25% de sujetos con puntuaciones más altas, y el 25% con puntuaciones más bajas en un test de actitudes conformado por ítems tipo Likert con 5 categorías de respuesta. ¿Podemos decir que el elemento discrimina de manera estadísticamente significativa?

	Sujeto	**Puntuación**
25% superior	20	10
	2	9
	13	7
25% inferior	3	4
	5	5
	8	2

4. En la tabla adjunta aparecen las respuestas de 200 sujetos a las tres alternativas de respuesta (A, B, C) de un ítem de un test, de las que la opción B es la correcta. Se sabe que la media del test, una vez descontada las puntuaciones correspondientes al ítem, es de 12 puntos. También se presentan las medias obtenidas en el test por los sujetos que respondieron a cada alternativa.

	A	***B****	***C***
50% superior	31	58	19
50% inferior	30	30	32
Media test	9	14	12

4.1. Calcular el índice de dificultad del ítem.

4.2. Sabiendo que la varianza corregida de las puntuaciones empíricas en el test es 9, calcular el índice de discriminación del ítem. Justifica la elección del índice utilizado.

4.3. Comentar los resultados obtenidos y la calidad del conjunto de alternativas.

5. Para investigar la posibilidad de sesgo en contra de los sujetos introvertidos en un ítem de un test de selección de personal, se llevó a cabo un análisis del funcionamiento diferencial de los ítems. Por ese motivo, se formaron dos grupos, uno de extrovertidos (GR), y otro de introvertidos (GF) a partir de las puntuaciones que se habían obtenido en otro test de perso-

nalidad previamente validado. En la siguiente tabla se muestra el número de respuestas adecuadas (*A*) e inadecuadas (*I*) de los extrovertidos e introvertidos en función de los niveles de adecuación al perfil del puesto establecidos por el test de selección de personal que van de *1* (nada adecuado) hasta *5* (muy adecuado). Analizar si existe FDI.

	Extrovertidos (GR)		Introvertidos (GF)	
Nivel de adecuación	A	I	A	I
1	3	6	1	10
2	11	36	6	45
3	59	28	15	66
4	78	10	43	32
5	80	9	46	29

6. Ejercicios conceptuales

1. El índice de dificultad sin corregir de un ítem dicotómico coincide con el promedio de aciertos en el ítem.
2. El poder discriminativo de un ítem se puede estimar mediante el coeficiente de correlación biserial puntual entre las puntuaciones de los sujetos en el ítem y las obtenidas en un criterio externo al test.
3. El índice de validez de un ítem se define como la correlación entre las puntuaciones obtenidas en el ítem y las puntuaciones en el test.
4. A medida que los ítems seleccionados para conformar un test sean más fiables más alta será su validez.
5. Un distractor de un ítem discrimina adecuadamente cuando los sujetos con bajo nivel en el test tienden a acertar el ítem.
6. Al aumentar el número de alternativas de respuesta de los ítems se reduce la probabilidad de acertar por azar.
7. El método de Mantel-Haenszel sólo informa sobre cuál es el grupo perjudicado por el ítem con sesgo, pero no sobre posibles motivos.
8. Seleccionar ítems con máxima fiabilidad y validez garantiza que las propiedades métricas del test sean óptimas.

9. El coeficiente *phi* se utiliza para estudiar la relación de un ítem con un criterio que sólo adopta dos posibles valores.
10. La dimensionalidad del test es independiente de la discriminación de los ítems.

10. SOLUCIONES A LOS EJERCICIOS DE AUTOEVALUACIÓN

1.

1.1.

La dificultad en el grupo total es:

$$ID = \frac{A}{N} = \frac{6}{10} = 0,60$$

$$ID_c = p - \frac{q}{k-1} = 0,60 - \frac{0,40}{3-1} = 0,40$$

En los 5 sujetos menos competentes es:

$$ID = \frac{2}{5} = 0,40$$

$$ID_c = 0,40 - \frac{0,60}{3-1} = 0,10$$

Mientras que en los 5 sujetos más competentes es:

$$ID = \frac{4}{5} = 0,80$$

$$ID_c = 0,80 - \frac{0,20}{3-1} = 0,70$$

Las conclusiones que debemos obtener son:

a) en primer lugar que la dificultad de los ítems depende claramente del nivel de competencia de la muestra de sujetos. De esta forma, para los sujetos menos hábiles el ítem ha tenido una dificultad media-alta; mientras que para los más hábiles ha sido extremadamente fácil: *b)* cuando utilizamos el ID_c la dificultad siempre aumenta porque contrarresta el efecto de acertar por azar; y esta corrección es mayor en la muestra de sujetos menos hábiles porque se supone que es más verosímil que respondan sin conocer el contenido del ítem y por tanto acierten por azar.

1.2. Para calcular el índice de discriminación, una primera aproximación es restar la proporción de aciertos entre el grupo más competente y el menos:

$$D = p_s - p_i = 0,80 - 0,40 = 0,40$$

De acuerdo con la tabla 8.5 podemos concluir que el ítem discrimina adecuadamente. Si consideramos conjuntamente la dificultad obtenida en toda la muestra y la discriminación encontrada, tendríamos que considerar que se trata de un buen ítem.

2.

2.1. La discriminación obtenida mediante el índice *D* es:

$$D = p_s - p_i = 0,70 - 0,32 = 0,38$$

2.2. La correlación biserial-puntual viene dada por:

$$r_{bp} = \frac{\overline{X}_A - \overline{X}_T}{S_x}\sqrt{\frac{p}{q}}$$

Para estimarla, primero preparamos adecuadamente la tabla de respuestas, destacando que la alternativa *b* es la correcta:

	Ítems			Total	
Sujetos	1 (a)	2 (b*)	3 (c)	X	$(X-i_2)$
A	1	1	1	3	2
B	1	0	1	2	2
C	0	1	1	2	1
D	0	0	0	0	0
E	1	1	1	3	2

Los sujetos que ha acertado el ítem 2 son el *A, C* y *E*. Su media en el test es:

$$\overline{X}_A = \frac{2+1+2}{3} = 1,67$$

La media total del test es:

$$\overline{X}_{T-i} = \overline{X}_{x-i} = \frac{2+2+1+0+2}{5} = 1,40$$

La desviación típica de las puntuaciones del test:

$$S^2_{x-i} = \frac{2^2+2^2+1^2+0^2+2^2}{5} - (1,4)^2 = 0,64$$
$$S_{x-i} = \sqrt{0,64} = 0,8$$

La proporción de sujetos que ha acertado el ítem 2 es 3/5 = 0,60; mientras la de sujetos que lo han fallado es 2/5 = 0,4.

$$r_{bp} = \frac{\overline{X}_A - \overline{X}_{T-i}}{S_{x-i}}\sqrt{\frac{p}{q}} = \frac{1,67-1,4}{0,8}\sqrt{\frac{0,6}{0,4}} = 0,41$$

La correlación biserial viene dada por:

$$r_b = \frac{\overline{X}_A - \overline{X}_{T-i}}{S_{x-i}}\frac{p}{y}$$

Si buscamos en las tablas el valor de *y* encontramos que vale 0,3863:

$$r_b = \frac{\overline{X}_A - \overline{X}_{T-i}}{S_{x-i}}\frac{p}{y} = \frac{1,67-1,4}{0,8}\frac{0,6}{0,3863} = 0,52$$

La relación entre la r_{bp} y la r_b:

$$r_{bp} = r_b\frac{y}{\sqrt{pq}} = 0,52\frac{0,3863}{\sqrt{0,6\cdot 0,4}} = 0,41$$

Se observa que la correlación biserial siempre es mayor que la biserial-puntual. En cualquier caso, el ítem presenta un buen índice de discriminación.

2.3. Para calcular la discriminación del distractor *c* del ítem 1 procedemos de la misma manera.

	Ítems			Total	
Sujetos	1 (a)	2 (b)	3 (c)	X	$(X-i_1)$
A	1	1	1	3	2
B	1	0	1	2	1
C	0 (c)	1	1	2	2
D	0	0	0	0	0
E	1	1	1	3	2

El sujeto que ha elegido la opción *c* en el ítem 1 es el *C*, luego su media es:

$$\overline{X}_A = \frac{2}{1} = 2$$

La media total del test es:

$$\overline{X}_{T-i} = \frac{2+1+2+0+2}{5} = 1,40$$

La desviación típica de las puntuaciones del test:

$$S^2_{x-i} = \frac{2^2+1^2+2^2+0^2+2^2}{5} - (1,4)^2 = 0,64$$
$$S_{x-i} = \sqrt{0,64} = 0,80$$

La proporción de sujetos que ha acertado el ítem 1 es 3/5 = 0,60; mientras la de sujetos que lo ha fallado es 2/5 = 0,40.

$$r_{bp} = \frac{\overline{X}_A - \overline{X}_{T-i}}{S_{x-i}}\sqrt{\frac{p}{q}} = \frac{2-1,4}{0,8}\sqrt{\frac{0,6}{0,4}} = 0.91$$

Por tanto, se trata de un distractor malo porque, precisamente el sujeto que lo ha seleccionado ha obtenido una puntuación media alta en el test.

3.

$$T = \frac{\bar{X}_{sj} - \bar{X}_{ij}}{\sqrt{\frac{(n_s-1)S_{sj}^2 + (n_i-1)S_{ij}^2}{n_s+n_i-2}\left(\frac{1}{n_s}+\frac{1}{n_i}\right)}} = \frac{8,67-3,67}{\sqrt{\frac{(3-1)1,56+(3-1)1,56}{3+3-2}\left(\frac{1}{3}+\frac{1}{3}\right)}} = 4,9$$

El valor empírico obtenido es de 4,9. Para un NC del 95% el valor teórico que encontramos en las tablas para 4 grados de libertad (3 + 3 – 2) es de 2,13. Dado que el valor empírico obtenido en los datos de nuestra muestra es mayor que el teórico, deberíamos rechazar la hipótesis nula que establece que las medias para ambos grupos son estadísticamente iguales, es decir, el ítem discrimina adecuadamente.

4.

4.1.

$$ID = \frac{A - \frac{E}{K-1}}{N} = \frac{88 - \frac{112}{2}}{200} = 0,16$$

4.2.

$$r_{bp} = \frac{\bar{X}_A - \bar{X}_{T-i}}{S_{x-i}}\sqrt{\frac{p}{q}} = \frac{14-12}{3}\sqrt{\frac{0,44}{0,56}} = 0,59$$

4.3. En función de la proporción de respuestas a los distractores, parece que el ítem funciona adecuadamente. Es decir, no hay ninguno que sea manifiestamente falso. Los sujetos menos competentes responden a los distractores aproximadamente en la misma proporción; mientras que los más hábiles identifican claramente la opción correcta y no hay ningún distractor que atraiga sus respuestas en una proporción elevada.

5.

Nivel 1

	C	I	
GR	3	6	
GF	1	10	
			20

Nivel 2

	C	I	
GR	11	36	
GF	6	45	
			98

Nivel 3

	C	I	
GR	59	28	
GF	15	66	
			168

Nivel 4

	C	I	
GR	78	10	
GF	43	32	
			163

Nivel 5

	C	I	
GR	80	9	
GF	46	29	
			164

Niveles de aptitud	$(A_I \times D_I)/N_I$	$(B_I \times C_I)/N_I$
Nivel I	(3 × 10)/20 = 1,5	(6 × 1)/20 = 0,3
Nivel II	(11 × 45)/98 = 5,05	(36 × 6)/98 = 2,20
Nivel III	(59 × 56)/168 = 23,18	(28 × 15)/168 = 2,5
Nivel IV	(78 × 32)/163 = 15,31	(10 × 43)/163 = 2,64
Nivel V	(80 × 29)/164 = 14,15	(9 × 46)/164 = 2,52
Total	59,19	10,16

$$\alpha_{MH} = \frac{\sum_{i=1}^{n} \frac{A_i D_i}{N_i}}{\sum_{i=1}^{n} \frac{B_i C_i}{N_i}} = \frac{59,19}{10,16} = 5,82$$

Dado que $\alpha_{MH} > 1$, el ítem favorece claramente al grupo de referencia, en este caso al grupo de los extrovertidos, tal como se había sospechado.

6. Ejercicios conceptuales

1. Verdadero.
2. Falso.

   Para calcular el poder discriminativo, hemos de considerar únicamente las puntuaciones del test del que el ítem forma parte. Cuando el criterio es externo, dicha correlación se podría interpretar mejor como un indicador de validez del ítem.
3. Falso.

   Se trataría de la correlación de las puntuaciones obtenidas en el ítem con las obtenidas en un criterio externo al test.
4. Falso.

   Se debe a una paradoja que ocurre en la Teoría Clásica de los Tests dado que la validez se puede representar a partir de la discriminación de los ítems, de su fiabilidad y de la dificultad, se da el caso de que el sumatorio de los *IF* (índices de fiabilidad de los ítems) es el denominador de la ecuación que relaciona dichos conceptos, y por tanto cuanto más elevado es el denominador más pequeña se hace la validez.
5. Falso.

   Es justo lo contrario, un distractor funciona adecuadamente cuando los sujetos que tienden a seleccionarlo son los que han puntuado bajo en el test. La función de una opción incorrecta es precisamente atraer la atención de los sujetos menos competentes.
6. Verdadero.

   Se reduce la probabilidad de acertar porque un sujeto poco hábil tendrá más opciones incorrectas entre las que elegir.
7. Verdadero.

   Una vez detectado funcionamiento diferencial mediante Mantel-Haenszel, el estudio de las causas del FDI se circunscribe al sesgo.

8. Falso.

   Vimos anteriormente que seleccionar ítems con máxima fiabilidad redunda en una reducción de la validez. Por tanto habrá que buscar un equilibrio entre ambas, y aun cuando obtengamos un test con estadísticos óptimos puede darse el caso de que no sea adecuado para nuestros objetivos, que carezca de validez de contenido, o aparente, por ejemplo.
9. Verdadero.
10. Falso.

    La discriminación está muy influida por el número de conceptos implicados en la obtención de las puntuaciones del test. Sólo tiene sentido estimar la discriminación de los ítems dentro de la escala a la que pertenecen, por ello cuando tras un Análisis Factorial obtengamos varias dimensiones en un test, la discriminación de cada ítem hemos de hallarla dentro de su dimensión y no considerando únicamente la puntuación global del test porque entonces estaríamos subestimándola.

11. BIBLIOGRAFÍA BÁSICA

Barbero, I. (2007). *Psicometría II: Métodos de elaboración de escalas*. Madrid: UNED.

Capítulo VII: La técnica de Likert para la medida de las actitudes.

Martínez-Arias, M.T., Hernández, M.J. y Hernández, M.V. (2006). *Psicometría*. Madrid: Alianza Editorial.

Capítulo 3: La Teoría Clásica de los Tests II: puntuaciones, análisis de elementos, consideraciones finales.

Muñiz, J. (2003). *Teoría Clásica de los Tests*. Madrid, Pirámide.

Capítulo 4: Análisis de ítems.

Muñiz, J., Martinez, R., Moreno, R., Fidalgo, A. y Cueto, E. (2005). *Análisis de los ítems*. Madrid: La Muralla.

Parte III

APLICACIÓN DE LOS INSTRUMENTOS Y EVALUACIÓN DE LOS SUJETOS

TEMA 9

ASIGNACIÓN, TRANSFORMACIÓN Y EQUIPARACIÓN DE LAS PUNTUACIONES

Enrique Vila Abad

SUMARIO

1. ORIENTACIONES DIDÁCTICAS

A lo largo de los temas anteriores se ha abordado el problema de la construcción de los instrumentos de medición psicológica y de la evaluación de su calidad métrica. Disponemos, por lo tanto, de un instrumento que nos va a permitir llevar a cabo la medición de la variable de interés. Queda, no obstante, una parte muy importante que es la siguiente: una vez elaborada la prueba definitiva hay que aplicarla, asignar puntuaciones a cada sujeto y dotar de significado a esas puntuaciones para poderlas interpretar. Esta última etapa es la que estudiaremos en este tema, puesto que la forma de aplicación del test y la de asignación de puntuaciones a los sujetos se estudió en los temas 2 y 3 cuando se abordó el problema de la construcción y aplicación de la prueba piloto.

La interpretación de las puntuaciones comienza justificando la necesidad de transformar las puntuaciones empíricas, que se han obtenido al aplicar un test a un grupo de sujetos, para conseguir una información fácilmente comprensible tanto para los sujetos a los que se ha aplicado el test, como para todas aquellas personas que estén interesadas en su significado y, una vez hecha esta justificación, se presentan los procedimientos más utilizados para llevar a cabo esa transformación.

Dentro de las transformaciones lineales de las puntuaciones, hacemos alusión a las escalas típicas y a las escalas típicas derivadas. Entre las transformaciones no lineales veremos las tres más utilizadas: el rango de percentiles, las escalas típicas normalizadas y las escalas normalizadas derivadas. Se incluyen también las normas cronológicas.

La segunda cuestión que abordamos en la exposición del tema es el concepto de equiparación de puntuaciones. Comenzamos con una breve descripción del concepto de equiparación para, a continuación, presentar los diseños y los métodos mas utilizados. Dentro de estos métodos hacemos referencia al método de la media, al método lineal y al método equipercentil.

Al estudiar el tema se recomienda profundizar en los siguientes puntos básicos:

— El objetivo que se persigue con el proceso de transformación de las puntuaciones.

— Tipos básicos de normas.
— Transformaciones lineales y no lineales.
— Normas cronológicas.
— Concepto de equiparación.
— Diseños de equiparación.
— Métodos de equiparación.

2. NECESIDAD DE TRANSFORMACIÓN DE LAS PUNTUACIONES PARA SU INTERPRETACIÓN

Cuando aplicamos un test, o un conjunto de tests, a un sujeto, lo corregimos y le asignamos una puntuación, ésta representa una descripción cuantitativa del rasgo que estamos evaluando. Ahora bien, ¿cómo se interpreta esa puntuación? ¿qué significado tiene? Supongamos que aplicamos un test de comprensión lectora a un sujeto, y éste obtiene 60 puntos. El primer paso sería interpretar dicha puntuación. La cuestión es, cómo interpretarla y saber, si 60 puntos implican mucha o poca comprensión lectora. Si nuestro interés se centra, solamente, en conocer la posición relativa de este sujeto respecto al resto de sus compañeros de clase (grupo normativo), la simple ordenación de los sujetos según su puntuación, sería suficiente para obtener información respecto a si su capacidad de comprensión lectora es mayor o menor que la de sus compañeros. A la escala resultante de asignar a los sujetos una puntuación, se la suele denominar, escala primaria (Petersen y col., 1989). Sin embargo, en la mayoría de las situaciones reales, las cosas no resultan tan sencillas como en el ejemplo que acabamos de ver. A veces, aplicamos varios tests a un mismo sujeto y las puntuaciones obtenidas en cada uno de ellos pueden venir en escalas distintas con lo cuál es difícil poder compararlas; o bien, a partir de la puntuación obtenida por el sujeto hemos de tomar la decisión de si es apto o no para alguna cosa determinada. En estos casos, la interpretación de los resultados se hace más compleja y surge la necesidad de poder contar con procedimientos que nos permitan dar un significado a las puntuaciones obtenidas.

Los dos procedimientos de interpretación propuestos son: la interpretación normativa y la interpretación criterial.

En la interpretación referida a la norma, o normativa, se compara la puntuación obtenida por un sujeto en un test con las obtenidas, en el mismo test, por un grupo de referencia o *grupo normativo*. A las puntuaciones obtenidas por los sujetos que constituyen el grupo normativo, así como a las transformaciones que se hagan de dichas puntuaciones, se las denomina *normas*. El conjunto de todas las normas constituye el baremo del test.

En la interpretación referida al criterio, que surge en los años cincuenta a raíz del auge del enfoque conductista, el interés central, tal y como se ha expuesto en los temas precedentes, no estriba en definir la posición de un sujeto respecto de su grupo de referencia, sino que se basa en determinar el grado de dominio que un sujeto tiene sobre un criterio preestablecido. Para ello, se suele tomar una puntuación de corte, que permita clasificar a los sujetos en dos grupos: los que dominan el criterio definido y los que no lo dominan. Como se puede observar, el referente ya no es un grupo normativo sino un criterio previamente establecido.

Veamos un ejemplo en el que se combinan ambas interpretaciones:

Supongamos que una empresa desea promocionar a un determinado puesto de trabajo a varios corredores de bolsa. Para ello, les aplica un test compuesto por 70 ítems de elección múltiple, con una sola respuesta correcta y puntuados de forma dicotómica con un 1 si el sujeto responde el ítem correctamente y un 0 si lo hace de forma incorrecta. Uno de los empleados obtiene 40 puntos en dicha prueba. ¿Podríamos decir que el rendimiento de este sujeto sería el adecuado para el nuevo puesto?, ¿debería realizar un cursillo intensivo de formación y adecuación al nuevo puesto antes de ser promocionado? Si nos fijamos solamente en la puntuación obtenida en el test, poco podemos decir, salvo que de las 70 preguntas el empleado ha contestado correctamente 40. No sabremos si su rendimiento es el adecuado, o si debería o no realizar el cursillo de formación.

Para contestar a la primera pregunta es necesario seleccionar una muestra representativa de la población de sujetos que ocupan dicho puesto (grupo normativo), aplicarles la prueba de evaluación diseñada y, finalmente, determinar la distribución de frecuencias de las puntuaciones obtenidas en el test por los sujetos que forman la muestra; el siguiente paso, sería ver dónde se sitúa nuestro sujeto en dicha distribución, y si está por encima o por debajo del rendimiento medio obtenido por el grupo normativo. Si está por encima podríamos decir que el sujeto es adecuado al puesto.

Para contestar a la segunda pregunta, tendríamos que establecer un criterio que definiese cuándo un sujeto tiene el nivel necesario para acceder al puesto de trabajo porque ha superado un criterio y cuándo deberá seguir un cursillo de formación. Para ello, una vez definido éste, compararíamos la puntuación del sujeto con la puntuación crítica del criterio (punto de corte). Si la puntuación obtenida por el sujeto está por debajo del punto de corte el sujeto debería realizar el cursillo, y si la puntuación del sujeto está por encima no necesitaría realizarlo.

La puntuación del sujeto es la misma en ambas situaciones, 40 puntos, sin embargo, la interpretación que debemos darle para contestar a las dos cuestiones planteadas es muy distinta. En el primer caso, hemos llevado a cabo una interpretación referida a la norma al comparar la puntuación del sujeto con la obtenida por un grupo normativo externo y, en el segundo caso, hemos llevado a cabo una interpretación referida al criterio al establecer una puntuación de corte que delimita si un sujeto tiene o no que realizar el cursillo.

3. TRANSFORMACIÓN DE LAS PUNTUACIONES EN LOS TESTS REFERIDOS A NORMAS

Dado que se trata de una interpretación normativa, es necesario seleccionar de la población objeto de estudio una muestra representativa a la que se aplica el test (o los tests) y sobre esa muestra se obtienen todas las normas. Una vez establecidas estas normas, se puede comparar la puntuación obtenida por un sujeto perteneciente a la misma población para saber cual es su posición respecto a la del grupo normativo y, de esa manera, poder interpretar la puntuación que ha obtenido.

A partir de las puntuaciones directas de los sujetos que forman el grupo normativo se pueden obtener otras escalas, mediante una serie de transformaciones, que permitan una mejor interpretación de las mismas. Estas transformaciones pueden ser de dos tipos: *transformaciones lineales y transformaciones no lineales*. Dentro de las transformaciones lineales se van a presentar *la escala de puntuaciones típicas* y *la escala de puntuaciones típicas derivadas*. En cuanto a las transformaciones no lineales, se presentan los *rangos percentiles, las escalas típicas normalizadas* y *las escalas de puntuaciones derivadas normalizadas*.

3.1. Transformaciones lineales

3.1.1. Escalas típicas

Una primera transformación lineal de las puntuaciones directas son las *puntuaciones típicas*. Éstas, se definen, como la diferencia entre la puntuación empírica directa obtenida por un sujeto en un test y la media del grupo de referencia, dividida por la desviación típica de este mismo grupo en el test.

$$Z_x = \frac{X - \bar{X}}{S_x} \qquad [9.1]$$

donde:

X = puntuación directa.

$\bar{X}$ = media de la muestra.

S_x = desviación típica de la muestra.

La puntuación típica nos indica el número de desviaciones típicas a las que se encuentra la puntuación de un sujeto respecto de la media del grupo normativo o de referencia. Supongamos que la media obtenida por una muestra de sujetos en un test es igual a 9, su desviación típica 4 y que un sujeto obtiene una puntuación típica igual a 2. Eso quiere decir que la puntuación directa que ha obtenido el sujeto está a dos desviaciones típicas por encima de la media del grupo. Teniendo en cuenta que la desviación típica es igual a 4, la puntuación del sujeto estará a 8 puntos de la media; por lo tanto será igual a 9 + 8 = 17 puntos.

EJEMPLO:

Hemos aplicado un test de razonamiento a una muestra de 400 sujetos. Sabiendo que la media y la desviación típica obtenidas fueron: $\bar{X} = 18$ y $S_x = 3$, calcular la puntuación típica de dos sujetos cuyas puntuaciones directas en el test fueron, respectivamente, 16 y 21.

$$Z_1 = \frac{X - \bar{X}}{S_x} = \frac{16 - 18}{3} = \frac{-2}{3} = -0{,}67$$

$$Z_2 = \frac{X - \bar{X}}{S_x} = \frac{21 - 18}{3} = \frac{3}{3} = 1$$

El primer sujeto se encuentra a 0,67 desviaciones típicas por debajo de la media del grupo puesto que su puntuación típica es negativa y el segundo sujeto se encuentra a una desviación típica por encima de la media del grupo.

La escala de puntuaciones típicas tiene de media 0 y desviación típica 1. Asimismo, la distribución de puntuaciones típicas de una variable normal suele oscilar de –3 a +3, lo que implica la existencia de valores negativos y decimales. Una forma de evitar este inconveniente es el empleo de las *escalas típicas derivadas*.

3.1.2. Escalas típicas derivadas

Como acabamos de señalar, una forma de evitar el tener que trabajar con puntuaciones negativas o con decimales, consiste en el empleo de *escalas típicas derivadas*. Las escalas típicas derivadas, son transformaciones lineales de las escales típicas. Esta transformación consiste, esencialmente, en multiplicar la puntuación típica por una constante b, desviación típica de la nueva escala, y sumarle otra constante a, la media en la escala resultante. La transformación se puede expresar como:

$$Y = a + bZ_x \qquad [9.2]$$

donde:

Y = puntuación típica derivada.

a = media de las puntuaciones en la nueva escala.

b = desviación típica de las puntuaciones en la nueva escala.

Z_x = puntuación típica en la escala original.

Si bien existen diversas posibles transformaciones, las más utilizadas suelen ser la escala D y la escala T.

— ***Escala D***: $D = 50 + 20Z_x$

Se trata de una escala en la que la media es igual a 50 y la desviación típica es igual a 20. Para el ejemplo anterior tenemos:

$$Z_1 = -0{,}67 \rightarrow D = 50 + 20\,(-0{,}67) = 50 + (-13{,}4) = 36{,}6 \approx 37$$
$$Z_2 = 1 \rightarrow D = 50 + 20(1) = 70$$

— ***Escala T***: $T = 50 + 10Z_x$

En esta escala la media es igual a 50 y la desviación típica es igual a 10. Fue desarrollada por McCall (1939), con la finalidad de reflejar las puntuaciones de niños en tests de habilidad mental.

Para el ejemplo anterior tenemos:

$$Z_1 = -0{,}67 \rightarrow T = 50 + 10\,(-0{,}67) = 50 + (-6{,}7) = 43{,}3 \approx 43$$
$$Z_2 = 1 \rightarrow T = 50 + 10(1) = 60$$

Si bien el empleo de las escalas típicas derivadas resuelve el problema de tener que trabajar con valores negativos o con decimales, ya que cuando se obtienen valores decimales se deben redondear al valor entero más próximo, sigue persistiendo un problema también común a la escala típica: la aplicación de un test a distintas muestras de sujetos dará lugar, seguramente, a valores distintos tanto de la media como de la desviación típica y, en algunos casos, las distribuciones de las puntuaciones de los sujetos no serán siempre iguales. Una distribución puede ser asimétrica positiva y otra asimétrica negativa. De producirse este hecho, tendremos que tener cuidado a la hora de comparar la puntuación de un sujeto, con respecto a una muestra concreta, ya que los tipos de escalas que acabamos de ver, solamente representan una transformación lineal de la escala, pero no de la forma de la distribución. Una forma de resolver este problema, es el empleo de las *escalas típicas normalizadas*.

3.2. Transformaciones no lineales

3.2.1. Rango de percentiles

Se define el percentil como aquella puntuación del test que deja por debajo de sí un determinado porcentaje de casos del grupo normativo. Si decimos que la puntuación 40 equivale al percentil 90 queremos decir que esa puntuación deja por debajo al 90% de los sujetos de la muestra, o que es superior a la del 90% de los sujetos. El percentil nos proporciona una idea de la posición de un determinado sujeto dentro del grupo normativo. Los percentiles constituyen una escala ordinal.

Para calcular los percentiles aplicamos la siguiente expresión:

$$P_x \text{ ó } C_x = \frac{100}{N}\left(f_b + \frac{f_d}{I}(X_c - L_i)\right) = f_{ac}\frac{100}{N} \qquad [9.3]$$

donde:

P_x ó C_x = porcentaje de sujetos que obtienen una puntuación inferior a la puntuación directa X.

N = número de sujetos de la muestra.

f_b = frecuencia absoluta acumulada bajo del intervalo crítico = f_a.

f_d = frecuencia absoluta dentro del intervalo crítico = f_x.

I = amplitud de los intervalos.

X_c = puntuación del test correspondiente al centil C_x.

L_i = límite inferior del intervalo crítico.

f_{ac} = frecuencia acumulada al punto medio del intervalo donde se encuentra X_c.

Nota: Se asume que dentro del intervalo los sujetos se reparten homogéneamente de manera que si existen 10 sujetos en un determinado intervalo 5 de ellos quedarían por debajo del punto medio y otros 5 por encima.

EJEMPLO:

A continuación aparecen las puntuaciones obtenidas por un grupo de sujetos en una prueba de ortografía:

8, 6, 5, 7, 8, 9, 4, 6, 3, 6, 9, 4, 2, 10, 6, 7, 5, 1, 2, 2, 5, 3, 7, 4, 5

Si un sujeto obtiene 8 puntos en dicha prueba, ¿qué percentil representa esa puntuación?

En primer lugar, ordenamos las puntuaciones de menor a mayor, y calculamos la distribución de frecuencias y frecuencias acumuladas.

X	1	2	3	4	5	6	7	8	9	10
f_d	1	3	2	3	4	4	3	2	2	1
f_b	1	4	6	9	13	17	20	22	24	25

A continuación aplicamos la ecuación 9.3:

$$C_x = \frac{100}{N}\left(f_b + \frac{f_d}{I}(X_c - L_i)\right) = \frac{100}{25}\left(20 + \frac{2}{1}(8 - 7{,}5)\right) = 4(20 + 2 \cdot 0{,}5) = 84$$

o bien :

$$C_x = f_{ac}\frac{100}{N} = (21)4 = 84$$

Hay que tener en cuenta que en el intervalo que va desde 7,5 a 8,5 hay 2 sujetos, que el punto medio es 8 y que, por lo tanto, por debajo del punto medio queda un sujeto en ese intervalo; si a ese sujeto le añadimos todos los que hay en los intervalos inferiores (20) hasta el punto medio del intervalo habrá 21 sujetos que son los que aparecen en la fórmula.

Un sujeto que ha obtenido una puntuación de 8 puntos deja por debajo al 84% de los sujetos de la muestra, por lo tanto la puntuación de 8 representa el percentil 84.

Si queremos saber la puntuación que le corresponde a un sujeto que supera al X% de los sujetos de la muestra, simplemente despejamos de la expresión anterior el término X_c.

$$X_c = L_i + \left(\frac{N \cdot C_x}{100} - f_b\right)\frac{I}{f_d} \qquad [9.4]$$

EJEMPLO:

Con los datos del ejemplo anterior, queremos saber la puntuación de un sujeto que deja por debajo al 60% de los sujetos de la muestra. Es decir, la puntuación que corresponde al percentil 60.

$$X_c = L_i + \left(\frac{N \cdot C_x}{100} - f_b\right)\frac{I}{f_d} = 5{,}5 + \left(\frac{25 \cdot 60}{100} - 13\right)\frac{1}{4} = 5{,}5 + (15 - 13) \cdot 0{,}25 = 6$$

Un sujeto que obtiene una puntuación de 6 puntos, deja por debajo al 60% de los sujetos de la muestra. Por lo tanto la puntuación de 6 en el test representa el percentil 60.

Si calculamos la puntuación del sujeto que deja por debajo al 84%, tendremos:

$$7{,}5 + \left(\frac{25 \cdot 84}{100} - 20\right)\frac{1}{2} = 75 + 1 \cdot 0 \cdot 5 = 8$$

que es, lógicamente, la puntuación de la que partíamos en el primer ejemplo.

Dada su facilidad de interpretación, los percentiles son una de las puntuaciones de mayor uso en el campo de la psicología a la hora de presentar los resultados obtenidos por un sujeto en un test. Decir, como hemos visto en el ejemplo anterior, que un sujeto ocupa el percentil 84, equivale a decir que deja por debajo al 84% de los sujetos de la muestra. Esta escala también presenta la ventaja de que podemos comparar las puntuaciones de un mismo sujeto en tests distintos puesto que su significado es el mismo independientemente del test aplicado y de la forma de la distribución de frecuencias. Si Pablo obtiene un percentil 70 en tres tests, uno de aptitud numérica, otro de rendimiento académico y otro de fluidez verbal, el significado es idéntico para las tres pruebas. Es decir, en los tres casos supera al 70% de los sujetos del grupo de referencia.

Los percentiles también nos permiten comparar las puntuaciones de sujetos distintos en un mismo test. Supongamos que la puntuación que obtuvo Pablo en el test de fluidez verbal es 35, y que en esa misma prueba, Jaime obtiene una puntuación de 20. Los resultados indican aparentemente, que Pablo presenta un grado de fluidez verbal mejor que el de Jaime. Pero, ¿qué sucedería si ambos sujetos pertenecieran a grupos de edad distintos?, entonces se deberían comparar con las puntuaciones obtenidas por sus respectivos grupos normativos. Supongamos, que Pablo pertenece a una muestra de niños de 13 años, con un percentil 70, y que Jaime pertenece a una muestra de niños de 9 años con un percentil 80. En este caso, podemos decir que Jaime presenta un grado de fluidez verbal superior al de Pablo, en relación a su grupo normativo, aún teniendo en cuenta que la puntuación empírica obtenida en el test es inferior.

3.2.2. Escalas típicas normalizadas

Las puntuaciones típicas normalizadas se obtienen a partir de los percentiles, y se definen, como la puntuación típica que le corresponde a una puntuación empírica obtenida por un sujeto en un test en una distribución normal. Al emplear estas puntuaciones estamos asumiendo que la distribución

de las puntuaciones es una distribución normal o, en caso de que esto no ocurra, se fuerza y se modifica la forma de la distribución de manera que se ajuste a una distribución normal. Esto implica la necesidad de ser cautelosos a la hora de interpretar los resultados ya que si la distribución de las puntuaciones se alejara mucho de una distribución normal se podrían estar falseando los datos.

Para obtener estas puntuaciones debemos partir, como ya hemos adelantado, de los percentiles, y mediante la tabla de la curva normal, se busca el valor de la puntuación típica Z_n que les corresponde. Si utilizamos los datos del ejemplo anterior, primero calculamos los percentiles correspondientes a las puntuaciones directas obtenidas y, a continuación, se buscan en las tablas de la curva normal las puntuaciones típicas normalizadas.

En la tabla adjunta la primera fila corresponde a las puntuaciones directas obtenidas por los sujetos, estas puntuaciones directas representan el punto medio de una distribución de puntuaciones en la que la amplitud del intervalo es la unidad. Así la puntuación directa 9 equivale al punto medio del intervalo que incluye todos los valores que van desde 8,5 a 9,5, siendo 8,5 el límite inferior y 9,5 el límite superior del intervalo. La segunda y tercera filas corresponden a la distribución de frecuencias y las puntuaciones típicas respectivamente que se obtienen de la forma que se indica a continuación. En la cuarta columna se recogen las frecuencias acumuladas hasta el punto medio del intervalo, para obtener estas frecuencias hay que asumir que los sujetos incluidos en un intervalo se distribuyen homogéneamente de manera que hay el mismo número de sujetos por encima y por debajo del punto medio; entonces, supongamos que en el primer intervalo que hay 1 sujeto, para efectuar los cálculos diremos que quedaría 0,5 por encima y 0,5 por debajo. En el siguiente intervalo (puntuación 2) hay 3 sujetos, entonces habría 1,5 por encima del punto medio y 1,5 por debajo; luego por debajo de la puntuación 2 tendríamos a todos los sujetos que estaban en el primer intervalo (1) más la mitad de los que están en el segundo intervalo (1,5), tendríamos 2,5 sujetos. De esta manera iríamos construyendo la cuarta columna. En la quinta se han obtenido los percentiles correspondientes a los puntos medios de los intervalos y, finalmente, en la sexta columna se incluyen las puntuaciones típicas normalizadas, que son las que se obtienen acudiendo a la tabla de la distribución normal y buscando la puntuación típica correspondiente. Si la distribución de las puntuaciones de nuestro ejemplo se hubieran ajustado a una distribución normal estas puntuaciones serían iguales a las puntuaciones típicas incluidas en la tercera columna; en nuestro caso se observa que esto no es así.

X	F_a	Z_X	Frecuencias Acumuladas al Punto Medio	Percentil	Típica Normalizada
10	1	1,95	24,5	98	2,05
9	2	1,53	23	92	1,39
8	2	1,11	21	84	0,99
7	3	0,69	18,5	74	0,64
6	4	0,27	15	60	0,25
5	4	−0,15	11	44	−0,15
4	3	−0,57	7,5	30	−0,52
3	2	−0,99	5	20	−0,84
2	3	−1,41	2,5	10	−1,28
1	1	−1,83	0,5	2	−2,05

A modo de ejemplo, veamos como se obtienen los valores de la puntuación típica, el percentil y la puntuación títpica normalizada para el caso de una puntuación empírica directa $X = 10$.

$$\bar{X} = 5,36 \quad S_x = 2,38$$

$$Z_1 = \frac{X - \bar{X}}{S_x} = \frac{10 - 5,36}{2,38} = \frac{4,64}{2,38} = 1,95$$

$$P_x = \frac{100}{N}\left(f_b + \frac{f_d}{I}(X_c - L_i)\right) = \frac{100}{25}\left(24 + \frac{1}{1}(10 - 9,5)\right) = 4(24 + 1 \cdot 0,5) = 98$$

Para calcular la puntuación típica normalizada, buscamos en la tabla de distribución normal (incluida al final del libro) el valor correspondiente al percentil 98; es decir, la puntuación típica que deja por debajo el 98% de la distribución de puntuaciones. A dicho valor le corresponde una puntuación típica normalizada de 2,05.

Este proceso es el que se seguirá con el resto de las puntuaciones directas.

En el caso de que las puntuaciones se distribuyeran según la curva normal, las puntuaciones típicas y las típicas normalizadas coincidirían tal y como hemos dicho anteriormente. En este caso no sería necesario llevar a cabo el proceso de normalización. Asimismo, si la distribución de las puntuaciones se aleja demasiado de una distribución normal de puntuaciones, el proceso de normalización no sería conveniente ya que estaríamos falseando los datos, forzando las puntuaciones a una distribución irreal.

3.2.3. Escalas normalizadas derivadas

Al igual que sucedía con las escalas típicas, las escalas típicas normalizadas presentan el inconveniente de los valores negativos y decimales, lo cual puede resultar incómodo para trabajar, y hacer mas difícil la interpretación de los resultados para personal no especializado. Estos inconvenientes se pueden resolver, mediante la transformación de las puntuaciones típicas normalizadas a *puntuaciones derivadas normalizadas.*

La escala normalizada derivada más utilizada es la escala de *estaninos o eneatipos*. Se utilizó por primera vez durante la Segunda Guerra Mundial por el ejército de los Estados Unidos. La escala de estaninos consiste en una escala de valores enteros y positivos de 9 unidades, del 1 al 9. Esta escala derivada tiene de media 5 y desviación típica 2.

$$E = 5 + 2(Z_n) \qquad [9.5]$$

EJEMPLO:

Calcular el estanino correspondiente a las puntuaciones típicas normalizadas

$$Z_{n1} = 0{,}25 \text{ y } Z_{n2} = 0{,}64$$
$$E_1 = 5 + 2(Z_{n1}) = 5 + 2(0{,}25) = 5{,}5 \approx 6$$
$$E_2 = 5 + 2(Z_{n2}) = 5 + 2(0{,}64) = 6{,}28 \approx 7$$

En la siguiente tabla podemos observar la equivalencia que existe entre la escala de estaninos, porcentajes de la distribución normal y los percentiles.

Estaninos	1	2	3	4	5	6	7	8	9
Dist. Normal	4%	7%	12%	17%	20%	17%	12%	7%	4%
Percentiles	4	5–11	12–23	24–40	41–60	61–77	78–89	90–96	>96
Punt. Típicas	–2	–1,5	–1	–0,5	0	0,5	1	1,5	2

Es decir, el estanino o eneatipo 1 incluiría el 4% inferior de los valores de la distribución, el 2 el 7% siguiente y así sucesivamente. Para saber qué percentiles se incluirían en cada eneatipo, bastaría ir acumulando los porcentajes correspondientes a cada eneatipo; así el percentil 4 correspondería al eneatipo 1, al eneatipo 2 le corresponden los percentiles del 5 al 11 (4 + 7), al eneatipo 3 los percentiles del 12 al 23 (4 + 7 + 12) y así sucesivamente.

Esto hay que tenerlo en cuenta a la hora de calcular el eneatipo que le corresponde a un sujeto. Por ejemplo, si al aplicar la fórmula correspondiente se obtiene un valor de 2,2 ese sujeto ya estaría situado por encima del 11% de la distribución y, por lo tanto se le debería asignar el eneatipo 3 que incluiría desde el 11% inferior hasta el 23% (sería el 12% siguiente).

Esta escala presenta el inconveniente de que al incluir en el mismo eneatipo a sujetos con distintas puntuaciones, se pierde bastante información. Supongamos que en un test de aptitud, dos sujetos obtienen la puntuación de 6 y 7 puntos respectivamente. Supongamos además, que a la puntuación obtenida por el primer sujeto le corresponde el percentil 65 y, a la puntuación obtenida por el segundo sujeto el percentil 74. Como se puede observar, la diferencia entre un percentil y otro es notoria; sin embargo, a ambas puntuaciones les correspondería el estanino o eneatipo 6.

3.3. Normas cronológicas

Las normas cronológicas constituyen otro tipo de transformación de las puntuaciones directas obtenidas por un grupo de sujetos en un test. La interpretación de la puntuación obtenida por un sujeto en un test, se lleva a cabo con relación a su edad y con la puntuación media obtenida por los sujetos de su edad.

Para Crocker y Algina (1986), este tipo de escalas no son muy recomendables por los inconvenientes que presentan. No siempre es posible la comparación de las puntuaciones de un mismo sujeto en áreas distintas, puesto que a las mismas puntuaciones de edad pueden corresponderles rangos percentiles diferentes y, consiguientemente, tener significados distintos. En segundo lugar, el significado de un año de edad mental no es constante con el desarrollo evolutivo del niño. A medida que aumenta la edad cronológica, la distancia entre un año y el siguiente disminuye, con lo que se dificulta su interpretación. Consideremos lo que ocurre con el desarrollo intelectual. Durante la infancia se produce un desarrollo rápido y constante que va decreciendo a medida que llegamos a la adolescencia. Las diferencias, por ejemplo, en razonamiento son mayores entre los 8 y 9 años de edad que entre los 15 y los 16 años.

Las normas cronológicas más utilizadas son la *edad mental* y el *cociente intelectual.* Las escalas de *edad mental* fueron propuestas, en principio, por Alfred Binet y, posteriormente, por las investigaciones de Binet-Simon. En el proceso de construcción de este tipo de escalas se deben seleccionar, en primer lugar, muestras de niños correspondientes a los distintos rangos de edad contemplados en el test. En segundo lugar, se aplica el test a los niños de cada rango de edad, y se calcula la puntuación media del test para cada uno de los rangos de edad. En tercer lugar, se construye una tabla en la que se asigna a cada edad la puntuación media correspondiente en el test. Supongamos, por ejemplo, que en un test de razonamiento abstracto los niños de 9 años obtienen una puntuación media de 25 puntos. Si aplicamos dicho test a un niño y éste obtiene 25 puntos, la asignaremos la edad mental de 9 años, independientemente de su edad cronológica.

Debido a algunas de las razones expuestas, este tipo de escalas, dan lugar a interpretaciones equívocas, por lo que su utilización ha caído en desuso hoy en día.

Para la obtención del *cociente intelectual*, se calcula la edad mental del sujeto y se divide por su edad cronológica, multiplicando por 100 el valor obtenido.

$$CI = \frac{EM}{EC} \cdot 100 \qquad [9.6]$$

donde:

CI = cociente intelectual.

EM = edad mental.

EC = edad cronológica.

De esta ecuación se puede deducir que el cociente intelectual será igual a 100 para todos los sujetos en los que el valor de la edad mental y la edad cronológica coincida, siendo así para todas las edades. El cociente intelectual será menor de 100, cuando exista un nivel de desarrollo intelectual más bajo que el promedio de su grupo, y será mayor de 100, cuando exista un desarrollo intelectual más alto que el promedio de su grupo. Esta escala es poco recomendable debido a los inconvenientes que presenta. El cociente intelectual es poco discriminativo para los adultos debido a que la edad mental medida por los tests se estabiliza a partir de una determinada edad cronológica con lo que se produce el efecto de techo. Otro inconveniente es que las distintas distribuciones de cocientes intelectuales para distintas edades, no presentan la misma desviación típica. Esto implica que el mismo cociente intelectual no proporciona la misma posición relativa en las distribuciones de distintas edades.

4. EQUIPARACIÓN DE PUNTUACIONES

Las puntuaciones que obtiene un sujeto en un test proporcionan una información de considerable valor ya que, en muchas situaciones, son decisivas a la hora de tomar decisiones. En ocasiones, estas puntuaciones pueden servir para ayudar a un sujeto a tomar la decisión de qué estudios seguir, o si puede ser apto o no para una determinada tarea. En otras ocasiones, estas puntuaciones pueden ser decisivas, para determinar la admisión de un estudiante a cierta universidad o carrera, o para una empresa a la hora de seleccionar a un grupo de profesionales. En cualquier caso, sea cual sea la decisión que se vaya a tomar, lo más importante es que la información que nos proporcionen esas puntuaciones sean lo más precisas posible. Supongamos, por ejemplo, que un sujeto realiza por segunda vez un examen de admisión para una determinada empresa, y que obtiene una puntuación superior a la obtenida la primera vez que realizó dicha prueba. En

principio podemos pensar que la diferencia de puntuación entre ambas aplicaciones se puede deber a que dicho sujeto se ha esforzado más en la segunda prueba. También podríamos pensar, que en ambas ocasiones se le ha aplicado la misma prueba y que, por lo tanto, el hecho de obtener una puntuación más alta en la segunda ocasión, se debe a que recuerda algunas de las preguntas que se le habían formulado la primera vez. Afortunadamente, en estas situaciones se suelen emplear formas distintas y el segundo efecto no se suele dar.

Supongamos otra posible situación. Esta misma empresa anuncia una convocatoria para cubrir una serie de puestos de trabajo y, dado que el número de sujetos que se presentan a la convocatoria es muy elevado, decide realizar diferentes pruebas en días distintos. Una vez que tienen lugar dichas pruebas, vemos que Juan, quien realizó la prueba el primer día, obtiene una puntuación más alta que Pedro, que realizó la prueba el segundo día. Las diferencias encontradas pueden ser debidas, a que la preparación de Juan es superior a la de Pedro; pero, puede ser que la diferencia se deba a que la primera prueba era más sencilla que la segunda, en cuyo caso Juan estaría jugando con una clara ventaja. El proceso de equiparación puede resolver estos problemas.

Definimos la equiparación de las puntuaciones de dos o más tests, como:

> El proceso mediante el cual se establece una correspondencia entre las puntuaciones de dichos tests, de tal manera que sea indistinto el empleo de uno u otro, puesto que las puntuaciones de cualquiera de ellos se podrán expresar en términos de las del otro test (Kolen y Brennan, 1995; Martínez, 1995; Muñiz, 1998).

Si el proceso de equiparación entre las puntuaciones de Juan y Pedro se ha llevado a cabo correctamente, podremos conocer si las diferencias encontradas son atribuibles a una mayor sencillez de la prueba del primer día o a una mayor preparación por parte de Juan.

Para establecer la equiparación entre tests, hay dos cuestiones fundamentales: que los tests midan el mismo constructo psicológico y que lo hagan con la misma fiabilidad. Estas condiciones son necesarias si queremos equiparar correctamente las puntuaciones de tests distintos.

Los pasos a seguir para llevar a cabo el proceso de equiparación son:

- Definir el propósito de la equiparación.
- Construir formas diferentes del test.
- Elegir un diseño para la recogida de datos.
- Recogida de datos.
- Determinar el método a emplear para equiparar las puntuaciones.
- Evaluar los resultados obtenidos.

A continuación se describen tanto los diseños como los métodos de equiparación más utilizados, y que se refieren a lo que ha venido a denominarse como *equiparación horizontal*, es decir, equiparación entre las puntuaciones obtenidas en tests que a priori se han intentado construir con

la misma dificultad. Si la equiparación se lleva a cabo entre las puntuaciones obtenidas en tests que midiendo el mismo rasgo tienen una dificultad distinta se denomina *equiparación vertical*. Una situación típica de este tipo de equiparación se plantea cuando se quieren establecer comparaciones entre competencias que se incrementan con la edad, utilizando tests de diferente dificultad en cada edad (Muñiz, 1998).

4.1. Diseños de equiparación

Cuando se lleva a cabo un estudio de equiparación, es conveniente que el número de sujetos que se vaya a utilizar sea representativo de la población a la que va a ir destinado el test. Presentamos a continuación los tres diseños más utilizados: de un solo grupo, de grupos equivalentes y de grupos no equivalentes con ítems comunes.

4.1.1. Diseño de un solo grupo

En los diseños de un solo grupo se administran las dos formas del test, cuyas puntuaciones se desean equiparar, al mismo grupo de sujetos. Las dos formas del test deben medir la misma característica objeto de estudio y presentar el mismo grado de dificultad. Este diseño presenta un inconveniente que debemos tener en cuenta. Supongamos las dos formas de un test X e Y. Si aplicamos en primer lugar la Forma X, y a continuación la Forma Y, nos podríamos encontrar con que las posibles diferencias entre las puntuaciones obtenidas por los sujetos en una forma y otra fueran debidas al cansancio (si es que la Forma Y se aplica a continuación de la Forma X), o también podría estar incidiendo el efecto del orden de presentación de ambas formas, con lo que la Forma aplicada en segundo lugar podría dar la sensación de ser mas fácil. Por ello, si aplicamos este diseño, se debe asumir que el valor de las puntuaciones obtenidas por los sujetos en la segunda Forma del test, no están afectadas por habérseles aplicado con anterioridad una primera Forma.

Debido a que no siempre estamos en condiciones de asegurar la inexistencia de estos efectos, es más aconsejable la utilización de una variante de este diseño: el *diseño de un solo grupo contrabalanceado*. Una de las formas de poder evitar los posibles efectos del orden de administración de las dos Formas del test es mediante el contrabalanceo. En este caso, dividimos a los sujetos en dos subgrupos incluyendo en cada uno un 50% de la muestra. A continuación se administra a ambos subgrupos las dos Formas del test en orden inverso, es decir, al primer subgrupo le aplicamos primero la Forma X y luego la Forma Y, y al segundo grupo le aplicamos primero la Forma Y y luego la Forma X. De esta manera, podemos asegurar que ambas Formas se verán afectadas por igual, por los efectos del orden de aplicación, la fatiga, etc.

4.1.2. Diseño de grupos equivalentes

En este diseño, se extraen de la población y de forma aleatoria dos muestras de sujetos, y a cada muestra se le aplica una Forma del test. Por lo tanto, cada sujeto responde solamente a una de las formas. Otra forma posible para obtener muestras aleatorias y equivalentes, puede ser alternar las Formas en cada grupo, de tal manera, que al primer sujeto se le entregue la Forma X, al segundo la Forma Y, al tercero la Forma X y así sucesivamente. Este diseño presenta la ventaja, al igual que sucede con el diseño de contrabalanceo, de que se evitan los efectos de fatiga, aprendizaje u orden de aplicación. También hay que destacar la importancia de que ambos grupos sean equivalentes en la aptitud que mide el test para evitar sesgos en el proceso de equiparación.

4.1.3. Diseño de grupos no equivalentes con ítems comunes

Al diseño de grupos no equivalentes con ítems comunes, también se le suele denominar *diseño de anclaje* y se puede considerar el diseño más utilizado a la hora de llevar a cabo la equiparación de las puntuaciones en distintos tests. Este diseño se asemeja al anterior, en que a cada una de las muestras de sujetos se le administra solamente una forma del test, la Forma X o la Forma Y. La diferencia estriba, en que ambas muestras no tienen porqué ser equivalentes entre sí y que, además, a ambas muestras se les aplica un test común (Z) que permite establecer las equivalencias entre los tests a equiparar. Consiguientemente, cada sujeto contesta un test diferente y un test común. A este test común que contestan ambos grupos se le conoce como *test de anclaje*.

Este diseño presenta dos posibles modalidades: el *test de anclaje interno* y el *test de anclaje externo* (Kolen y Brennan, 1995). En el primer caso, se utiliza un conjunto de ítems comunes a ambos tests y éstos aparecen intercalados con el resto de los ítems propios de las dos Formas X e Y, cuyas puntuaciones se quieren equiparar. Las puntuaciones obtenidas en los ítems comunes se incluyen en la puntuación total de los sujetos en el test. En el segundo caso, el test de anclaje externo, los ítems comunes aparecen formando un test independiente y las puntuaciones obtenidas por los sujetos en ese test no se utilizan en el cómputo de la puntuación total de los sujetos en las formas a equiparar. En el primer caso se habla de *ítems de anclaje* y en el segundo de *test de anclaje*. En ambos casos los ítems comunes deben de ser lo más parecidos posible a los de las dos formas aunque no sea una condición imprescindible (Lord, 1980).

Otra cuestión a tener en consideración, es el número de ítems comunes que se deben emplear (Angoff, 1984; Harris, 1993; Petersen y col., 1983; Wingersky y col. 1987). La experiencia sugiere, que el número de ítems a utilizar debería ser, al menos, el 20% de la longitud total de un test compuesto por 40 ítems, excepto en el caso en que un test esté formado por un número elevado de ítems, en cuyo caso la utilización de 30 ítems comunes puede resultar suficiente. También se debe tener en cuenta el mayor o menor grado de heterogeneidad del test.

Las diferencias que se pueden presentar entre las puntuaciones obtenidas en ambas formas pueden ser debidas a las diferencias entre ambos grupos de sujetos, o bien a las diferencias entre ambas formas. Veamos un ejemplo que nos permita ver la forma de poder observar, si las posibles diferencias son debidas a la primera causa o a la segunda.

EJEMPLO:

Supongamos que aplicamos las Formas X e Y de un test compuesto por 80 ítems, de los cuales 16 ítems son comunes a ambas formas, a dos grupos de sujetos. En la siguiente Tabla aparecen las medias obtenidas por ambos grupos, tanto en la Forma aplicada como en los ítems comunes.

Los valores de las medias obtenidos por ambos grupos en los ítems comunes, nos sugieren que el nivel de conocimiento en el grupo-2 es superior al del grupo-1. El grupo-2 contesta correctamente el 80% de los ítems comunes, mientras que el grupo-1 contesta correctamente el 60%. El grupo-2 contestó correctamente un 20% de ítems más que el grupo-1.

Grupo	Forma *X*	Forma *Y*	Ítems comunes
1	59	—	9 (60%)
2	—	70	12 (80%)

La segunda cuestión que nos planteamos es si las diferencias encontradas en las puntuaciones obtenidas por los sujetos son debidas a diferencias entre las dos Formas. Para responder a esta pregunta, nos podemos plantear cuál hubiera sido la puntuación media para el grupo-2, si le hubiéramos aplicado a este grupo la Forma X. El grupo-2 contestó correctamente un 20% de ítems comunes más que el grupo-1. Así pues, podríamos pensar que el grupo-2 contestará un 20% más de ítems en la Forma X (teniendo en cuenta que el test está compuesto por 80 ítems, el 20% sería 16 ítems) que el grupo-1. Consiguientemente, si utilizamos este razonamiento, su puntuación sería 59 + 16 = 75. El grupo-2 tiene una puntuación media en la Forma Y de 70 puntos, y su puntuación esperada en la Forma X es de 75 puntos, luego la Forma X, aparentemente, es más fácil que la Forma Y.

4.2. Métodos de equiparación

En el apartado anterior hemos presentado los diseños más frecuentemente utilizados a la hora de llevar a cabo un proceso de equiparación. A continuación, se presentan los métodos de equiparación más utilizados para la obtención de puntuaciones equivalentes a partir de tests distintos que evalúan el mismo rasgo psicológico.

4.2.1. Método de la media

En el método de la media se asume que las puntuaciones obtenidas por una muestra de sujetos en uno de los test difieren en una cuantía constante de las puntuaciones obtenidas por una muestra de sujetos en el otro test. En esencia, lo que se pretende con este método es hacer corresponder las medias de los tests a equiparar (Muñiz, 1998). Sean *X* e *Y* dos tests distintos, cuyas puntuaciones queremos equiparar. Para toda puntuación *X* podemos establecer que:

$$\boxed{X^* = Y = X - \bar{X} + \bar{Y}} \qquad [9.9]$$

donde:

X^* = puntuación del test *Y* equivalente a una del test *X*.

X = puntuación del test *X*.

$\bar{X}$ = media del test *X*.

$\bar{Y}$ = media del test *Y*.

Supongamos dos tests *X* e *Y* cuyas medias son, respectivamente, 65 y 70. Según el método de la media, tendríamos que sumarle a toda puntuación del test *X*, 5 puntos para poder equiparar las puntuaciones de ambos tests o, lo que es lo mismo, restarle 5 puntos a toda puntuación del test *Y*. Según esto, una puntuación de 60 puntos en el test *X* sería lo mismo que una puntuación de 65 puntos en el test *Y*.

Para $X = 60$

$$X^* = Y = X - \bar{X} + \bar{Y} = X - 65 + 70 = X + 5 = 60 + 5 = 65$$

4.2.2. Método lineal

Al contrario de lo que sucede en el método de la media, donde se supone que las diferencias entre las puntuaciones obtenidas por los sujetos en ambos tests es constante, en el método lineal las diferencias entre las puntuaciones pueden variar. Por ejemplo, las diferencias entre las puntuaciones bajas en el test pueden ser mayores que las diferencias encontradas entre las puntuaciones altas.

Este método se basa en la equiparación de aquellas puntuaciones directas que tienen la misma puntuación típica. Es decir, una determinada puntuación perteneciente a un test *Y*, es equivalente a una puntuación perteneciente a un test *X* si ambas puntuaciones tienen idéntica puntuación *Z*,

con lo que $Z_x = Z_y$ (Angoff, 1984; Kolen y Brennan, 1995; Suen, H, 1990). Por lo tanto, la transformación de las puntuaciones correspondientes al test *X* en puntuaciones *Y*, viene determinada por una transformación lineal que podemos expresar como:

$$\frac{X-\bar{X}}{S_x}=\frac{Y-\bar{Y}}{S_y} \quad \text{y despejando,} \quad X^*=Y=\left(\frac{S_y}{S_x}\right)(X-\bar{X})+\bar{Y} \qquad [9.10]$$

o bien:

$$X^* = a\,(X-b) + c$$

donde:

X^* = puntuación del test *Y* equivalente a una puntuación del test *X*.

S_y = desviación típica de las puntuaciones del test *Y*.

S_x = desviación típica de las puntuaciones del test *X*.

X = puntuación del test *X*.

$\bar{X}$ = b = media del test *X*.

$\bar{Y}$ = c = media del test *Y*.

$a=\frac{S_y}{S_x}$ = cociente entre las desviaciones típicas.

EJEMPLO:

Supongamos que se aplica a una muestra de sujetos un test de razonamiento numérico, siendo la media de las puntuaciones 38 y la desviación típica 5. A una segunda muestra le aplicamos un test *Y*, también de razonamiento numérico, siendo la media de las puntuaciones igual a 46, y la desviación típica 7. Las dos muestras han sido extraídas de la misma población y son muestras equivalentes. Deseamos saber qué puntuación en el test *Y* sería equivalente a la puntuación 40 obtenida por un sujeto en el test *X*.

$$X^*=Y=\left(\frac{S_y}{S_x}\right)(X-\bar{X})+\bar{Y}=\left(\frac{7}{5}\right)(40-38)+46=2,8+46=48,8$$

Este resultado indica que la puntuación de 48,8 puntos en el test *Y* es la que corresponde a una puntuación de 40 puntos en el test *X*.

En este ejemplo, se ha aplicado a cada grupo de sujetos una forma distinta del test, es decir, sería la situación del *diseño de grupos equivalentes*.

Si se hubiera utilizada un *diseño de un solo grupo*, en el que se deben administrar los dos tests, cuyas puntuaciones se desean equiparar, al mismo grupo de sujetos pero en orden inverso, la transformación lineal se expresaría de la siguiente manera:

$$X^*=Y=\sqrt{\left(\frac{S^2_{y1}+S^2_{y2}}{S^2_{x1}+S^2_{x2}}\right)}\left(X-\frac{\bar{X}_1+\bar{X}_2}{2}\right)+\frac{\bar{Y}_1+\bar{Y}_2}{2} \qquad [9.11]$$

donde:

El subíndice 1 hace referencia a los valores obtenidos en el subgrupo 1(subgrupo al que se le aplicó en primer lugar el test *X* y en segundo lugar el test *Y*). El subíndice 2 hace referencia a los valores obtenidos en el subgrupo 2 (subgrupo al que se le aplicó en primer lugar el test *Y* y en segundo lugar el test *X*).

X^* = puntuación del test *Y* equivalente a una puntuación del test *X*.

S_{y1} y S_{y2} = desviación típica de las puntuaciones del test *Y* aplicado al subgrupo 1 y 2.

S_{x1} y S_{x2} = desviación típica de las puntuaciones del test *X* aplicado al subgrupo 1 y 2.

X = puntuación del test *X*.

$\bar{X}_1$ y $\bar{X}_2$= media del test *X* aplicado al subgrupo 1 y 2.

$\bar{Y}_1$ y $\bar{Y}_2$= media del test *Y* aplicado al subgrupo 1 y 2.

EJEMPLO:

Supongamos que se selecciona de una población una muestra aleatoria y, una vez dividida en dos subgrupos equivalentes se aplica al primer grupo un test *X* de razonamiento numérico obteniéndose una media de 38 puntos y una desviación típica igual a 5 y un test *Y* también de razonamiento numérico cuya media fue 46 y la desviación típica igual a 7. A un segundo grupo le administramos los mismos tests, pero en orden inverso, obteniendo los siguientes resultados: la media de las puntuaciones en el test *Y* es igual a 44, y la desviación típica es igual a 6 y, la media de las puntuaciones en el test *X* es igual a 40 y la desviación típica es igual a 8. Deseamos saber qué puntuación en el test *Y* sería equivalente a la puntuación 37 obtenida por un sujeto en el test *X*.

$$X^*=Y=\sqrt{\left(\frac{S^2_{y1}+S^2_{y2}}{S^2_{x1}+S^2_{x2}}\right)}\left(X-\frac{\bar{X}_1+\bar{X}_2}{2}\right)+\frac{\bar{Y}_1+\bar{Y}_2}{2}=\sqrt{\left(\frac{7^2+6^2}{5^2+8^2}\right)}\left(37-\frac{38+40}{2}\right)+\frac{46+44}{2}=43,04$$

La puntuación del test *Y* que equivaldría a una puntuación de 37 en el test *X* sería la de 43,04 puntos.

En tercer lugar, podríamos haber utilizado un *diseño de anclaje* en el que se cuenta con dos grupos de sujetos y a cada grupo se les administra una forma diferente del test, y un test de anclaje (*Z*) que es común a ambos grupos. Como ya hemos dicho anteriormente, las diferencias entre las puntuaciones obtenidas por los sujetos pueden ser debidas a que los sujetos difieren en el rasgo que estemos estudiando, o bien a que los tests utilizados presenten niveles de dificultad distintos.

En este caso la transformación lineal quedaría expresada en los siguientes términos:

$$X^* = Y = \left(\frac{\sqrt{S_{y2}^2 + b_{yz2}^2 \left(S_z^2 - S_{z2}^2\right)}}{\sqrt{S_{x1}^2 + b_{xz1}^2 \left(S_z^2 - S_{z1}^2\right)}} \right) \left[X - \left(\overline{X}_1 + b_{xz1} \left(\overline{Z} - \overline{Z}_1 \right) \right) \right] + \left[\overline{Y}_2 + b_{yz2} \left(\overline{Z} - \overline{Z}_2 \right) \right] \qquad [9.12]$$

donde:

X^* = puntuación del test *Y* equivalente a una puntuación del test *X*.

S_{x1}^2 = varianza de las puntuaciones en el test *X*, aplicado al grupo 1.

b_{xz1}^2 = pendiente de la recta de regresión de *X* sobre *Z*, en el grupo 1:

$$b_{xz1} = r_{xz1} \frac{S_{x1}}{S_{z1}}$$

S_z^2 = varianza de las puntuaciones del test *Z*, calculada sobre los sujetos de los grupos 1 y 2.

S_{z1}^2 = varianza de las puntuaciones del test *Z*, calculada sobre los sujetos del grupo 1.

S_{y2}^2 = varianza de las puntuaciones en el test *Y*, aplicado en el grupo 2.

b_{yz2}^2 = pendiente de la recta de regresión de *Y* sobre *Z*, determinada en el grupo 2.

$$b_{yz2} = r_{yz2} \frac{S_{y2}}{S_{z2}}$$

S_{z2}^2 = varianza de las puntuaciones en el test *Z*, calculada sobre los sujetos del grupo 2.

X = puntuación del test *X*.

$\overline{X}_1$ = media de las puntuaciones en el test *X*, aplicado en el grupo 1.

$\overline{Z}$ = media de las puntuaciones en el test *Z*, calculada sobre los sujetos de los grupos 1 y 2.

$\overline{Z}_1$ = media de las puntuaciones en el test *Z*, calculada sobre los sujetos del grupo 1.

$\overline{Y}_2$ = media de las puntuaciones en el test *Y*, aplicado en el grupo 2.

$\overline{Z}_2$ = media de las puntuaciones en el test *Z*, calculada sobre los sujetos del grupo 2.

EJEMPLO:

Supongamos que se dispone de dos formas *X* e *Y* de un test de fluidez verbal compuesto por 100 ítems de elección múltiple, y un test de anclaje *Z* compuesto por 20 ítems, y se aplica cada forma del test a un grupo de sujetos junto con el test *Z*. En la siguiente tabla aparecen los datos correspondientes a las dos formas del test y al test de anclaje. Deseamos saber qué puntuación en el test *Y* sería equivalente a la puntuación 85 obtenida por un sujeto en el test *X*.

Test *X*	Test *Y*	Test *Z*
$S_{x1} = 11$	$S_{y2} = 12,5$	$S_z = 9,5$
$b_{xz1} = 0,80$	$b_{xz2} = 0,95$	$S_{z1} = 10$
$\overline{X}_1 = 74$	$\overline{Y}_2 = 79$	$S_{z2} = 11$
		$\overline{Z} = 15,5$
		$\overline{Z}_1 = 14$
		$\overline{Z}_2 = 17$

$$X^* = Y = \left[\frac{\sqrt{S_{y2}^2 + b_{yz2}^2 \left(S_z^2 - S_{z2}^2\right)}}{\sqrt{S_{x1}^2 + b_{xz1}^2 \left(S_z^2 - S_{z1}^2\right)}} \right] \left[X - \left(\overline{X}_1 + b_{xz1} (\overline{Z} - \overline{Z}_1) \right) \right] + \left(\overline{Y}_2 + b_{yz2} (\overline{Z} - \overline{Z}_2) \right)$$

$$\left[\frac{\sqrt{S_{y2}^2 + b_{yz2}^2 \left(S_z^2 - S_{z2}^2\right)}}{\sqrt{S_{x1}^2 + b_{xz1}^2 \left(S_z^2 - S_{z1}^2\right)}} \right] = \sqrt{\frac{12,5^2 + 0,95^2 \left(9,5^2 - 11^2\right)}{11^2 + 0,80^2 \left(9,5^2 - 10^2\right)}} = \sqrt{\frac{128,5}{114,8}} = 1,06$$

$$\left[\overline{X}_1 + b_{xz1} \left(\overline{Z} - \overline{Z}_1 \right) \right] = 74 + 0,80 (15,5 - 14) = 75,2$$

$$\left[\overline{Y}_2 + b_{yz2} \left(\overline{Z} - \overline{Z}_2 \right) \right] = 79 + 0,95 (15,5 - 17) = 77,6$$

$$X^* = 1,06 (85 - 75,2) + 77,6 = 87,98 \approx 88$$

La puntuación equivalente en el test *Y* de un sujeto que obtiene una puntuación de 85 en el test *X* es de 88 puntos.

4.2.3. Método equipercentil

El método equipercentil (Braun y Holland, 1982; Kolen, 1984; Martínez, 1995) es el método de equiparación más habitual, consiste en equiparar aquellas puntuaciones cuyos percentiles son iguales. Por ejemplo, supongamos que a un sujeto que obtiene una puntuación directa de 25 en un test *X* de Fluidez Verbal, le corresponde un percentil de 70 y, a un sujeto que obtiene una puntuación directa de 29 en un test *Y* de Fluidez Verbal, le corresponde también un percentil de 70. Entonces, podremos decir que una puntuación directa de 25 en el test *X* equivale a una puntuación de 29 en el test *Y*.

Según Crocker y Algina (1986), los pasos a seguir para llevar a cabo el proceso de equiparación percentil, se pueden resumir en los siguientes apartados:

— Tenemos dos tests *X* e *Y*, cuyas puntuaciones queremos equiparar. En primer lugar, calculamos en cada test las puntuaciones percentiles que corresponden a cada una de las puntuaciones de ambos tests. Para calcular dichas puntuaciones aplicamos la ecuación vista en el apartado 3.2.1.

$$P_x \text{ ó } C_x = \frac{100}{N}\left(f_b + \frac{f_d}{A}(X_c - L_i)\right) = f_{ac}\frac{100}{N} \qquad [9.13]$$

donde:

P_x ó C_x = porcentaje de sujetos que obtienen una puntuación inferior a la puntuación directa *X*.

N = número de sujetos de la muestra.

f_b = frecuencia absoluta acumulada bajo del intervalo crítico.

f_d = frecuencia absoluta dentro del intervalo crítico.

A = amplitud de los intervalos.

X_c = puntuación del test correspondiente al centil C_X.

L_i = límite inferior del intervalo crítico.

f_{ac} = frecuencia acumulada al punto medio del intervalo donde se encuentra X_c.

— En segundo lugar, representamos gráficamente las dos distribuciones de percentiles. Para ello, en el eje de abscisas ponemos las puntuaciones obtenidas por los sujetos en el test *X* y en el test *Y*. En el eje de ordenadas los rangos percentiles. A continuación, dibujamos la curva correspondiente a cada test.

— En tercer lugar, obtenemos las puntuaciones equivalentes en los dos tests *X* e *Y* a partir del gráfico anterior.

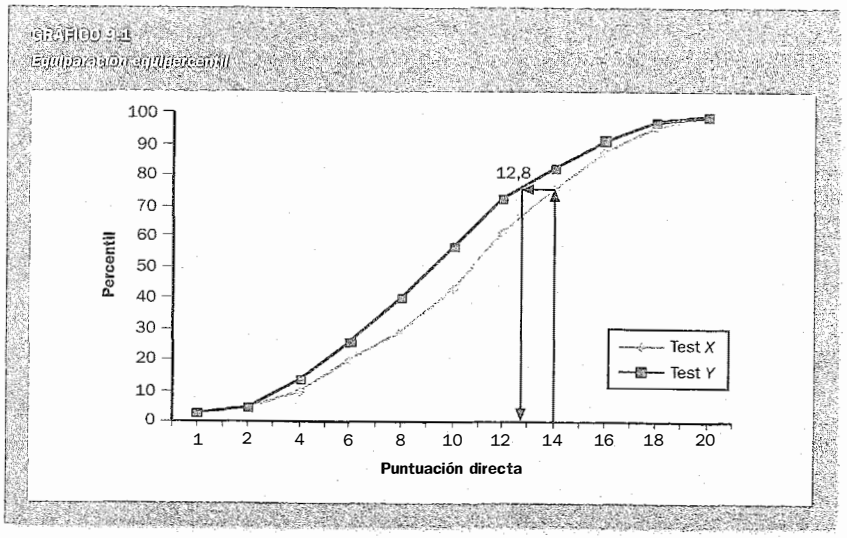
EJEMPLO:

En la tabla adjunta se presentan las puntuaciones percentiles correspondientes a un grupo de sujetos en dos formas (*X*, *Y*) de un test de razonamiento compuesto por 10 ítems.

Puntuación directa	Test X	Test Y
1	3	3
2	5	5
4	10	14
6	20	26
8	29	40
10	43	57
12	61	72
14	75	82
16	87	91
18	98	97
20	99	99

En el gráfico podemos observar como a una puntuación $X = 14$ le corresponde, aproximadamente, una puntuación equivalente $X^* = 12{,}8$. A partir de la puntuación 14, trazamos una línea perpendicular hasta cortar con la curva de distribución de percentiles del test *X*. En dicho, punto trazamos una línea perpendicular hasta cortar con la curva de distribución de percentiles del test *Y*. Trazamos una línea perpendicular hasta cortar con el eje de abscisas y determinamos la puntuación equipercentil equivalente, en este caso 12,8. Este proceso es el que se seguiría con el resto de las puntuaciones.

GRÁFICO 9.1
Equiparación por anclaje/percentil



En la siguiente tabla se presentan las puntuaciones X^* correspondientes a todas las puntuaciones. En la primera columna se presentan los rangos percentiles; en la segunda, las puntuaciones directas obtenidas en el test X; en la tercera, las puntuaciones equipercentiles equivalentes; y, en la cuarta, las puntuaciones equipercentiles equivalentes redondeadas.

Percentil	3	5	10	20	29	43	61	75	87	98	99
Punt. X	1	2	4	6	8	10	12	14	16	18	20
Punt. X^*	1	2	3,3	4,7	6,4	8,3	10,1	12,8	15,2	16,9	20
X^* redondeada	1	2	3	5	6	8	10	13	15	17	20

El procedimiento que acabamos de ver es idéntico para el diseño de un solo grupo y el diseño de grupos equivalentes. El diseño de anclaje presenta una mayor complejidad y el lector interesado puede consultar el texto de Angoff (1971).

5. ERROR TÍPICO DE EQUIPARACIÓN

El proceso de equiparación de puntuaciones no está libre de error aleatorio. Lord (1950), define el error típico de equiparación como la desviación típica de las puntuaciones transformadas a la escala Y, que se corresponden a un valor concreto de un test X.

$$\left(S_e = S_{(X^*/X)}\right)$$

Según Angoff (1984), el error típico de medida para las puntuaciones equiparadas se puede expresar de la siguiente manera:

Diseño de grupos equivalentes:

$$S_e = \sqrt{\frac{2S_y^2}{N_1 + N_2}(Z_x^2 + 2)} \qquad [9.14]$$

donde:

N_1 y N_2 = número de sujetos en ambas muestras.

Z_X = puntuación típica correspondiente al valor de X^*.

S_y^2 = varianza de las puntuaciones en el test Y:

$$Z_x^2 = \left[\frac{X^* - \bar{X}}{S_x}\right]^2$$

A medida que las puntuaciones equiparadas (X^*) se alejan de la media el error típico es mayor.

EJEMPLO:

Supongamos que aplicamos a una muestra de 50 sujetos un test X de percepción del color, donde la media de las puntuaciones en el test es igual a 20, y la desviación típica es igual a 4. A una segunda muestra, también de 50 sujetos, le aplicamos un test Y, también de percepción del color, donde la media de las puntuaciones es igual a 25, y la desviación típica es igual a 6. Las dos muestras han sido extraídas de la misma población y son muestras equivalentes. Deseamos saber qué puntuación en el test Y sería equivalente a la puntuación 30 obtenida por un sujeto en el test X y cuál es el error típico de equiparación cometido.

En primer lugar, calculamos la puntuación equiparable en el test Y.

$$X^* = Y = \left(\frac{S_y}{S_x}\right)(X - \bar{X}) + \bar{Y} = \left(\frac{6}{4}\right)(30 - 20) + 25 = 15 + 25 = 40$$

$$S_e = \sqrt{\frac{2S_y^2}{N_1 + N_2}(Z_x^2 + 2)} = \sqrt{\frac{2S_y^2}{N_1 + N_2}\left(\left(\frac{(X^* - \bar{X})}{S_x}\right)^2 + 2\right)} =$$

$$= \sqrt{\frac{2 \cdot 36}{100}\left(\left(\frac{40 - 20}{4}\right)^2 + 2\right)} = 4,41$$

Diseño de un solo grupo:

$$S_e = \sqrt{\frac{(S_y^2)(1 - r_{xy})\left(Z_x^2(1 + r_{xy}) + 2\right)}{N}} \qquad [9.15]$$

donde:

r_{xy} = correlación entre ambos tests.

S_y^2 = varianza de las puntuaciones obtenidas en el test *Y*: $S_y^2 = \frac{S_{y1}^2 + S_{y2}^2}{2}$

EJEMPLO:

Seleccionamos una muestra aleatoriamente de la población de 50 sujetos. Una vez dividida en dos subgrupos, aplicamos al primero un test *X* de aritmética, donde la media de las puntuaciones en el test es igual a 35, y la desviación típica es igual a 5; y un test *Y* también de aritmética donde la media de las puntuaciones es igual a 40, y la desviación típica es igual a 6. Al segundo subgrupo le administramos los mismos tests, pero en orden inverso, obteniendo los siguientes resultados: la media de las puntuaciones en el test *Y* es igual a 41, y la desviación típica es igual a 6 y, la media de las puntuaciones en el test *X* es igual a 38 y la desviación típica es igual a 7. La correlación entre ambos tests es igual a 0,80, y los valores totales del test: $\bar{X} = 36,5$, $\bar{Y} = 40,5$, $S_x = 6$.

Deseamos saber qué puntuación en el test Y sería equivalente a la puntuación 40 obtenida por un sujeto en el test *X* y cuál es el error típico de equiparación cometido.

En primer lugar, calculamos la puntuación equiparable en el test *Y*.

$$X^* = Y = \left(\frac{S_{y1} + S_{y2}}{S_{x1} + S_{x2}}\right)\left(X - \frac{\bar{X}_1 + \bar{X}_2}{2}\right) + \frac{\bar{Y}_1 + \bar{Y}_2}{2} = \left(\frac{6 + 6}{5 + 7}\right)\left(40 - \frac{35 + 38}{2}\right) + \frac{40 + 41}{2} = 44$$

$$S_e = \sqrt{\frac{(S_y^2)(1 - r_{xy})\left(Z_x^2(1 + r_{xy}) + 2\right)}{N}} = \sqrt{\frac{(S_y^2)(1 - r_{xy})\left(\left(\frac{(X^* - \bar{X})}{S_x}\right)^2(1 + r_{xy}) + 2\right)}{N}} =$$

$$= \sqrt{\frac{36(1 - 0,80)\left((1,56 \cdot 1,8) + 2\right)}{50}} = 0,83$$

Diseño de anclaje:

$$S_e = \sqrt{\frac{2S_y^2(1 - r^2)\left[Z_x^2(1 + r^2) + 2\right]}{N}} \qquad [9.16]$$

donde:

b_{xz1} = pendiente de la recta de regresión de *X* sobre *Z*, determinada en el grupo 1.

b_{yz2} = pendiente de la recta de regresión de *Y* sobre *Z*, determinada en el grupo 2.

EJEMPLO:

Utilizando los datos del ejemplo utilizado para el diseño de anclaje ($N = 100$):

Vimos que la puntuación equivalente en el test Y de un sujeto que en el test *X* obtuvo una puntuación de 85 es 88.

Test *X*	Test *Y*	Test *Z*
$S_{x1} = 11$	$S_{y2} = 12,5$	$S_z = 9,5$
$b_{xz1} = 0,80$	$b_{xz2} = 0,95$	$S_{z1} = 10$
$\bar{X}_1 = 74$	$\bar{Y}_2 = 79$	$S_{z2} = 11$
$X^* = 88$		$\bar{Z} = 15,5$
		$\bar{Z}_1 = 14$
		$\bar{Z}_2 = 17$

$$r = \frac{b_{xz1}}{S_x} = \frac{b_{yz2}}{S_y} = 0,072$$

$$Z_X = \frac{X^* - \overline{X}}{S_x} = \frac{88 - 74}{11} = 1,27$$

$$S_e = \sqrt{\frac{2S_y^2(1-r^2)\left(Z_x^2(1+r^2)+2\right)}{N}} = \sqrt{\frac{2 \cdot 156,25(1-0,005)\left(1,61(1+0,005)+2\right)}{100}} = 3,35$$

6. EL MANUAL DEL TEST

En los temas precedentes, y a lo largo de éste, hemos ido exponiendo aquellos aspectos psicométricos que son necesarios para la elaboración de un test. De todas estas consideraciones se podría concluir que un test se utiliza para obtener unas puntuaciones que hemos de interpretar y dotar de un cierto significado psicológico. Esto conlleva la necesidad, por parte del constructor de un test, de proporcionar una información a los usuarios del mismo de modo que puedan dar una significación adecuada a la puntuación obtenida por un sujeto en el test. Esta necesidad implica que el test incluya, además del propio test, el *manual del test*, que resulta imprescindible para llevar a cabo una óptima comprensión y aplicación del test. Como aspectos imprescindibles de dicho manual, se debe reflejar en qué consiste el test, las distintas fases de su construcción, para qué sirve y, las normas de aplicación y valoración. Todo ello se expuso en el tema 2.

El manual debe tener una finalidad práctica y, por lo tanto, no es necesario que se introduzca todo el material y consideraciones por las que pasó el autor. En caso necesario, podemos hacer referencia en el manual a otras posibles fuentes de información más extensas sobre el test.

Siguiendo a Yela (1984), en el manual deben figurar todos los datos que hacen del test un instrumento científico. Estos datos son susceptibles de ser agrupados en cuatro categorías: la especificación, la descripción, la justificación y las referencias bibliográficas.

— La especificación del test

Hace referencia a la denominación y clasificación del test. La clasificación la podemos subdividir en función del constructo psicológico que queremos evaluar, en la forma en que se presenta el material que empleamos en el test (impreso o manipulativo), o según el método de administración de la prueba (individual o colectiva).

— La descripción del test

Hace referencia a toda información relativa a los fines y forma de aplicación del test.

En primer lugar, podemos incluir una introducción donde se explique el objetivo del test, y sus principales características. También resultará útil, saber si el test guarda algún tipo de relación con otros tests similares. Por último, podemos incluir, de forma resumida, los antecedentes y desarrollo del test.

En segundo lugar, se especificará el campo de aplicación al que va dirigido el test. Incluiremos información sobre los aspectos psicológicos que se pretenden estudiar, áreas de aplicación a los que puede interesar de una manera especial y, otras aplicaciones que se hayan hecho de él así como los resultados obtenidos.

En tercer lugar, consideraremos la descripción detallada del material que incluye. Cabe hacer una diferenciación entre el material básico y el material auxiliar. En el primer caso, nos referimos al material de que consta el test (partes que lo componen, número de piezas de que consta, etc.). En el segundo caso, nos referimos a material auxiliar, como lápices, hojas de respuesta, cronómetros, etc.

En cuarto lugar, nos encontramos con un punto de máxima importancia: las instrucciones de aplicación. De su correcta aplicación dependerán en muchos casos las puntuaciones que obtenga un sujeto. Antes de comenzar, es muy conveniente dar una serie de pautas de carácter general, sobre atención, interés, comprensión, etc., de cada una de las tareas que se van a realizar, así como instrucciones específicas sobre su ejecución. Por último, se indicarán los tiempos exactos de los que se dispone para la ejecución de cada una de las partes del test.

En quinto lugar, incluiremos información respecto a la forma de puntuar. Se incluirán las plantillas con soluciones. En el caso de pruebas de carácter manipulativo se tendrá en cuenta la manera exacta de considerar una respuesta como acierto o error, y la forma de cronometrar con precisión. Se indicará, con ejemplos si fuera necesario, la forma de anotar las puntuaciones directas obtenidas.

— La Justificación

Con la justificación se incluyen los datos cuantitativos y experimentales que justifican el uso del test, y que permiten la valoración de sus resultados. Dentro de la justificación, se incluye toda la información relativa a la duración de la prueba, la fiabilidad, validez y tipificación del test.

— Referencias bibliográficas

Se incluirán todas aquellas referencias que contengan cualquier tipo de información referida al test.

A continuación se presentan algunas de las normas propuestas por la *American Psychological Association*, para la elaboración del manual de un test. Con ellas, se pretende resaltar algunos de

los aspectos que consideramos más interesantes, y que no constituyen, en modo alguno, todas las normas existentes. Para ello, se recomienda la consulta de editoriales de tests. El lector interesado puede encontrar listas en manuales de psicodiagnóstico o evaluación.

- En toda prueba debe llevarse a cabo una actualización periódica, y se desaconseja el empleo de pruebas que no se hayan actualizado en los últimos 15 años.
- Los manuales actualizados incluirán, además de los nuevos resultados obtenidos, los obtenidos por otros estudios y autores, y se reflejarán tanto los resultados positivos como los negativos.
- Si se hallase nueva información sobre el test que fuese contradictoria a la existente, se llevará a cabo una revisión y actualización del test lo antes posible.
- Cualquier revisión de un test implicará un nuevo análisis y tratamiento estadístico que aparecerá por separado en el manual.
- El manual debe incluir ejemplos sobre la interpretación de los datos y estadísticos del test. En estos casos, se harán constar los coeficientes y valores más significativos para aquellas situaciones que puedan considerarse complejas.
- En el manual se informará, de ser necesario, sobre la existencia de error sistemático.
- Se especificarán las distintas aplicaciones haciendo una diferenciación entre las de carácter práctico, de las de carácter de investigación.
- La redacción de las instrucciones y las normas de aplicación serán presentadas de tal forma que conlleven a reproducir siempre la misma situación. Además serán de fácil comprensión para los sujetos evaluados. En caso de que quien administra la prueba pueda introducir variaciones en las instrucciones, se hará constar en el manual.
- Cualquier información de carácter cuantitativo será presentada con la mayor precisión y claridad posible, añadiendo cuantos ejemplos sean necesarios para su adecuada interpretación.
- Para su correcta interpretación, es esencial que figure toda la bibliografía referente al test.
- Los criterios de puntuación han de estar perfectamente definidos, y deben incluir información acerca de posibles dudas, rectificaciones, comentarios, etc. Asimismo se incluirá información sobre las posibles alternativas en la corrección de las puntuaciones, y la posible necesidad de aplicar fórmulas de corrección del azar.
- En el manual se incluirá información de la fiabilidad y error de medida del test, así como la relativa a los ítems: dificultad, varianza, discriminación.
- Se harán constar los inconvenientes que representa la interpretación de resultados en pruebas que poseen una baja fiabilidad. Aparte de informar sobre las garantías de fiabilidad en las puntuaciones, se describirán los procedimientos y muestras a partir de los cuales se obtuvieron dichos resultados. Con respecto a las muestras, es conveniente tener una información sobre sus características personales y demográficas.

- La fiabilidad de pruebas de rendimiento académico, inteligencia y aptitudes, se calculará para cada grupo de edad y curso académico en el que vaya a ser aplicado. Si el test va a ser aplicado en grupos distintos, se calculará el coeficiente de fiabilidad en cada uno de éstos.
- Si el test consta de dos o más formas se proporcionará una breve descripción de las características estadísticas de cada una de ellas por separado. En este caso es interesante presentar las posibles semejanzas entre los ítems de cada una de las formas.
- Si aplicamos la técnica test-retest se incluirá el tiempo transcurrido entre una aplicación y otra, así como qué condiciones llevaron a establecer dicho intervalo.
- En pruebas que incluyan varios subtests correlacionados con el rango de puntuaciones globales, se incluirán tablas de equivalencia en las cuales se asigne para cada centil la puntuación en los distintos subtests.
- En tests compuestos de varios subtests, se presentará una matriz de correlaciones entre sus puntuaciones, así como los estadísticos descriptivos más significativos.
- En el manual se establecerá la estabilidad de las puntuaciones en el tiempo, y los factores que pueden afectar a dicha estabilidad. Para comprobar la estabilidad de las puntuaciones se utilizarán formas paralelas del mismo.
- El manual incluirá el período de caducidad en la validez de las puntuaciones del test.
- La información sobre la validez del test se referirá a los usos y aplicaciones concretas del instrumento.
- La validez de contenido del test, vendrá referida al sector del dominio que está reflejado en los ítems. El análisis del contenido y los criterios seguidos para la confección de los ítems no se debe confundir con los criterios externos de validación.
- Se describirá el proceso de selección y calidad de los criterios utilizados en el proceso de validación del test. Se incluirán todos los coeficientes de validez obtenidos con los criterios seleccionados.
- En situaciones en que se haya utilizado la validez predictiva, se hará referencia a la generalización de resultados entre muestras, distintas situaciones, etc.
- La homogeneidad de las conductas seleccionadas como criterio es un dato fundamental en la interpretación de su relación con el test. El manual incluirá el tiempo transcurrido entre la administración del test y la obtención de los datos del criterio. También constará la formación y preparación de los sujetos tanto en el momento de aplicar el test como en la obtención del criterio.
- La interpretación y valoración de los datos acerca de la validez ha de tener en cuenta las principales variables personales de los sujetos. Por ejemplo, en el caso de aplicar un test para

una selección de personal, se dará información respecto a las funciones, características y cometidos del puesto de trabajo para el cual se está utilizando el test.

- Se deben actualizar los valores de validez y comprobar los cambios que se producen en el tiempo.
- En tests de orientación escolar, se presentarán datos sobre la relación del test con la aptitud verbal de los sujetos. En tests de velocidad, se justificará la posible influencia de la rapidez en las puntuaciones obtenidas.
- La interpretación de las puntuaciones obtenidas en el test, así como la escala en que se expresan dichas puntuaciones ha de ser fácil de llevarse a cabo. Así mismo, se deben justificar las razones por las que se ha escogido una determinada escala. Si se producen revisiones posteriores de dichas escalas, se incluirán tablas de equivalencia entre las escalas originales y las revisadas.
- Los baremos, o conjunto de normas establecidas para la evaluación de los sujetos, que se presenten en el manual deben estar actualizados en todo momento y ser adecuados para futuras aplicaciones. Si los baremos se han obtenido a partir de muestras pequeñas y poco representativas, se advertirá en el manual de esta circunstancia y sus posibles implicaciones.
- Se dará información sobre los resultados de los distintos grupos empleados, teniendo en cuenta características de edad, sexo, nivel educativo, etc.

7. EJERCICIOS DE AUTOEVALUACIÓN

1. A un grupo de sujetos se les ha aplicado un test Razonamiento. La media de dicho grupo es 25 y desviación típica 8. Suponiendo que las puntuaciones se distribuyen según la curva normal, calcular: la puntuación típica, percentil y eneatipo que obtendría un sujeto que obtuvo en el test una puntuación empírica igual a 30.

2. Se ha aplicado un test de habilidades sociales a una muestra de 500 sujetos. Las puntuaciones obtenidas por los sujetos se distribuyen según la curva normal con media igual a 18 y desviación típica igual a 6. Calcular:

a. La puntuación típica, típica derivada de media 50 y desviación típica 20 y eneatipo que le corresponde a un sujeto que obtuvo en el test una puntuación directa igual a 24.

b. La puntuación directa de un sujeto que es superior al 75% de los sujetos de la muestra.

c. ¿Cuántos sujetos han obtenido puntuaciones inferiores a la media de la muestra?

3. Hemos aplicado a una muestra de 80 sujetos un test para evaluar su capacidad de comprensión lectora. Los datos obtenidos aparecen recogidos en la tabla adjunta:

X	f
16	5
14	15
12	40
10	17
8	3

Sabiendo que la distribución de las puntuaciones se ajusta a una distribución normal, calcular:

a. La puntuación centil (percentil) correspondiente a cada una de las puntuaciones directas

b. Las puntuaciones típicas y puntuaciones T de McCall.

c. Eneatipos

4. El equipo psicopedagógico de un colegio ha desarrollado dos formas (*X*,*Y*) de un test para evaluar la actitud de los profesores de primaria del colegio hacia los alumnos evaluados como hiperactivos. Para ello aplicaron el test *X* a 10 profesores y, a otros diez la forma *Y*. Ambos grupos se establecieron de forma aleatoria. Con los datos que se presentan a continuación, ¿cuáles serían las puntuaciones del test *Y* que equivaldrían a las del test *X*?

Forma X	Forma Y
50	31
41	38
42	42
51	39
37	41
53	46
50	34
54	42
48	37
53	52

5. La dirección de una empresa ha solicitado a su departamento de recursos humanos que evalúe la capacidad de gestión de sus empleados en las dos sucursales que posee. Puesto que no es posible llevar a cabo la evaluación de las dos sucursales a la vez, se han confeccionado dos tests distintos, de 40 preguntas cada uno. De las cuarenta preguntas, 10 son comunes a ambos tests y 30 diferentes. Las puntuaciones obtenidas por los cinco empleados de cada sucursal son:

Sucursal A		Sucursal B	
Ítems Comunes	Ítems diferentes	Ítems Comunes	Ítems diferentes
7	22	6	20
6	18	8	25
9	26	5	15
4	13	7	24
8	24	5	21

Calcular para cada empleado su calificación final en el test, de modo que las calificaciones de los cinco sujetos estén en la misma escala.

6. Preguntas conceptuales

A continuación se les presentan una serie de afirmaciones que deberá leer atentamente y decir si son correctas o incorrectas.

1. En los tests referidos a la norma, la puntuación obtenida por un sujeto se compara con un grupo normativo.
2. En los tests referidos al criterio se estudian, fundamentalmente, las diferencias existentes entre los sujetos.

3. Las escalas típicas derivadas son transformaciones lineales de las escalas típicas.
4. Los percentiles son transformaciones lineales de las puntuaciones directas.
5. El percentil es el porcentaje de sujetos que hay en la distribución del grupo normativo.
6. Las escalas típicas normalizadas se obtienen por transformación lineal de las escalas típicas derivadas.
7. La escala de estaninos es una escala de 9 unidades.
8. El cociente intelectual es igual a cien cuando el valor de la edad mental coincide con la edad cronológica.
9. Al diseño de equiparación de un solo grupo, también se le conoce con el nombre de diseño de anclaje.
10. En los diseños de equiparación de grupos equivalentes administramos las dos formas del test al mismo grupo de sujetos.
11. El método de la media supone que las diferencias entre las puntuaciones de dos tests es constante.
12. Si un sujeto ocupa el percentil 78, deja por debajo al 22% de los sujetos de la muestra.

8. SOLUCIONES A LOS EJERCICIOS DE AUTOEVALUACIÓN

1.

$\bar{X} = 25$

$S_x = 8$

$X = 30 \rightarrow Z_x = \dfrac{30-25}{8} = 0,62 \rightarrow P = 73,24 \approx 73$

$E = 5 + 2(0,625) = 6,25 \approx 6$

2.

a) $N = 500 \qquad \bar{X} = 18$

$S_x = 6$

$X = 24 \rightarrow Z_x = \dfrac{24-18}{6} = 1$

$PD = 50 + 20(1) = 70$

$E = 5 + 2(1) = 7$

Debido a que el enunciado establece que las puntuaciones obtenidas por los sujetos se distribuyen según la curva normal, no sería necesario llevar a cabo el proceso de normalización, siendo las puntuaciones típicas iguales a las puntuaciones típicas normalizadas.

b) En la tabla de la curva normal, la puntuación típica que deja por debajo al 75% de los sujetos es igual a $Z = 0,67$

$$Z_x = \frac{X - \bar{X}}{S_x};\ X = Z_x \cdot S_x + \bar{X} = 0,67 \cdot 6 + 18 = 22,02$$

c) Puesto que el enunciado establece una distribución normal de puntuaciones, podemos establecer que la puntuación $X = 18$ deja por debajo al 50% de los sujetos, es decir 250.

3.

X	f	fX	fX²
16	5	80	1280
14	15	210	2940
12	40	480	5760
10	17	170	1700
8	3	24	192
	80	964	11872

$\bar{X} = \dfrac{964}{80} = 12,05 \qquad S_x^2 = \dfrac{11872}{80} - (12,05)^2 = 148,4 - 145,2 = 3,2 \Rightarrow S_x = 1,79$

$Z_1 = \dfrac{8-12,05}{1,79} = -2,26 \Rightarrow P = 1,1 \approx 1$

$Z_2 = \dfrac{10-12,05}{1,79} = -1,14 \Rightarrow P = 12,71 \approx 13$

$Z_3 = \dfrac{12-12,05}{1,79} = -0,03 \Rightarrow P = 49,20 \approx 49$

$Z_4 = \dfrac{14-12,05}{1,79} = 1,09 \Rightarrow P = 86,21 \approx 86$

$Z_5 = \dfrac{16-12,05}{1,79} = 2,21 \Rightarrow P = 98,64 \approx 99$

b) Las puntuaciones típicas ya se han calculado en el apartado anterior

$Z_1 = -2,26 \qquad Z_2 = -1,14 \qquad Z_3 = -0,03 \qquad Z_4 = 1,09 \qquad Z_5 = 2,21$

$T = 50 + 10Z_n$

$T_1 = 50 + 10(-2,26) = 27,4 \approx 27 \qquad T_2 = 50 + 10(-1,14) = 38,6 \approx 39$

$T_3 = 50 + 10(-0,03) = 49,7 \approx 50 \qquad T_4 = 50 + 10(1,09) = 60,9 \approx 61$

$T_5 = 50 + 10(2,21) = 72,1 \approx 72$

c) $E = 5 + 2(Z_n)$

$E_1 = 5 + 2(-2,26) = 0,48 \approx 1 \qquad E_2 = 5 + 2(-1,14) = 2,72 \approx 3$

$E_3 = 5 + 2(-0,03) = 4,94 \approx 5 \qquad E_4 = 5 + 2(1,09) = 7,18 \approx 8$

$E_5 = 5 + 2(2,21) = 9,42 \approx 9$

4. Nos encontramos ante un diseño de grupos equivalentes. Por lo tanto, la ecuación de equiparación se define como:

$$X^* = Y = \left(\frac{S_y}{S_x}\right)(X - \bar{X}) + \bar{Y}$$

Forma X	Forma Y
50	31
41	38
42	42
51	39
37	41
53	46
50	34
54	42
48	37
53	52

$$S_x^2 = \frac{23253}{10} - (47,9)^2 = 2325,3 - 2294,4 = 30,9 \Rightarrow S_x = 5,56$$

$$S_y^2 = \frac{16480}{10} - (40,2)^2 = 1648 - 1616 = 32 \Rightarrow S_y = 5,65$$

$$X^* = \left(\frac{S_y}{S_x}\right)(X - \bar{X}) + \bar{Y} = 1,02(X - 47,9) + 40,2$$

Aplicando dicha ecuación a las puntuaciones de la forma *X* tenemos:

$$X^* = 1,02(X - 47,9) + 40,2 = 1,02(50 - 47,9) + 40,2 = 42,3$$
$$X^* = 1,02(X - 47,9) + 40,2 = 1,02(41 - 47,9) + 40,2 = 33,1$$
$$X^* = 1,02(X - 47,9) + 40,2 = 1,02(42 - 47,9) + 40,2 = 34,2$$
$$X^* = 1,02(X - 47,9) + 40,2 = 1,02(51 - 47,9) + 40,2 = 43,4$$

$$X^* = 1,02(X - 47,9) + 40,2 = 1,02(37 - 47,9) + 40,2 = 29,1$$
$$X^* = 1,02(X - 47,9) + 40,2 = 1,02(53 - 47,9) + 40,2 = 45,4$$
$$X^* = 1,02(X - 47,9) + 40,2 = 1,02(50 - 47,9) + 40,2 = 42,3$$
$$X^* = 1,02(X - 47,9) + 40,2 = 1,02(54 - 47,9) + 40,2 = 46,4$$
$$X^* = 1,02(X - 47,9) + 40,2 = 1,02(48 - 47,9) + 40,2 = 40,3$$
$$X^* = 1,02(X - 47,9) + 40,2 = 1,02(53 - 47,9) + 40,2 = 45,4$$

5.

Sucursal A		Sucursal B	
Ítems Comunes	Ítems diferentes	Ítems Comunes	Ítems diferentes
7	22	6	20
6	18	8	25
9	26	5	15
4	13	7	24
8	24	5	21

Se ha utilizado un diseño de anclaje. La ecuación de conversión es:

$$X^* = Y = \left[\frac{\sqrt{S_{y2}^2 + b_{yz2}^2\left(S_z^2 - S_{z2}^2\right)}}{\sqrt{S_{x1}^2 + b_{xz1}^2\left(S_z^2 - S_{z1}^2\right)}}\right]\left[X - \left(\bar{X}_1 + b_{xz1}(\bar{Z} - \bar{Z}_1)\right)\right] + \left(\bar{Y}_2 + b_{yz2}\left(\bar{Z} - \bar{Z}_2\right)\right.$$

Se calcula la media y varianza del grupo *A* y *B*:

$$A: S_{x1}^2 = 21,44; \quad \bar{X}_1 = 20,6; \quad S_{z1}^2 = 2,96; \quad \bar{Z}_1 = 6,8$$

$$B: S_{x2}^2 = 12,4; \quad \bar{X}_2 = 21; \quad S_{z2}^2 = 1,36; \quad \bar{Z}_2 = 6,2$$

Grupo total $(A + B)$: $S_z^2 = 2,25; \quad \bar{Z} = 6,5$

$$b_{xz1} = r_{xz1}\frac{S_{x1}}{S_{z1}} \qquad b_{xz1} = r_{xz1}\frac{S_{x1}}{S_{z1}} = 0,99\frac{4,63}{1,72} = 2,66$$

$$b_{yz2} = r_{yz2}\frac{S_{y2}}{S_{z2}} \qquad b_{yz2} = r_{yz2}\frac{S_{y2}}{S_{z2}} = 0,83\frac{3,52}{1,16} = 2,52$$

Z_1	X	XZ_1	Z_1^2	X^2	Z^2	Y	YZ_2	Z_1^2	Y^2
7	22	154	49	484	6	20	120	36	400
6	18	108	36	324	8	25	200	64	625
9	26	234	81	676	5	15	75	25	225
4	13	52	16	169	7	24	168	49	576
8	24	192	64	576	5	21	105	25	441
34	103	740	246	2229	31	105	668	199	2267

$$r_{XZ1} = \frac{N\cdot\sum XZ_1 - \sum X\sum Z_1}{\sqrt{\left[N\cdot\sum X^2 - (\sum X)^2\right]\left[N\cdot\sum Z_1^2 - (\sum Z_1)^2\right]}} =$$

$$= \frac{5\cdot 740 - 103\cdot 34}{\sqrt{(5\cdot 2229 - 103^2)(5\cdot 246 - 34^2)}} = \frac{3700 - 3502}{199,15} = 0,99$$

$$r_{YZ2} = \frac{N\cdot\sum YZ_2 - \sum Y\sum Z_2}{\sqrt{\left[N\cdot\sum Y^2 - (\sum Y)^2\right]\left[N\cdot\sum Z_2^2 - (\sum Z_2)^2\right]}} =$$

$$= \frac{5\cdot 668 - 105\cdot 31}{\sqrt{(5\cdot 2267 - 105^2)(5\cdot 199 - 31^2)}} = \frac{3340 - 3255}{102,66} = 0,83$$

$$X^* = Y = \left(\frac{\sqrt{12,4 + 6,15(2,25 - 1,36)}}{\sqrt{21,44 + 7,07(2,25 - 2,96)}}\right)$$

$$\left[X - (20,6 + 2,66(6,5 - 6,8))\right] + (21 + 2,52(6,5 - 6,2) =$$

$$X^* = 1,04\cdot(X - 19,80) + 21,76$$

X	13	18	22	24	26
X^*	15	20	24	26	28

6. Soluciones a las preguntas conceptuales

1. La afirmación es correcta.
2. La afirmación es falsa.

   En los tests referidos al criterio se intenta determinar el grado de dominio que un sujeto tiene sobre un criterio o materia determinada.
3. La afirmación es correcta.
4. La afirmación es falsa.

   Los percentiles son transformaciones no lineales.
5. La afirmación es falsa.

   Los percentiles son puntuaciones que dejan por debajo un determinado porcentaje de sujetos.
6. La afirmación es falsa.

   Se obtienen a partir de los percentiles.
7. La afirmación es correcta.
8. La afirmación es correcta.
9. La afirmación es falsa.

   El diseño de anclaje hace referencia al diseño de grupos no equivalentes a los que se aplican tests distintos con ítems comunes.
10. La afirmación es falsa.

    Se seleccionan aleatoriamente de una población dos muestras equivalentes y, a cada una de ellas, se le aplica una forma del test.
11. La afirmación es verdadera.
12. La afirmación es falsa.

    Dejará por debajo al 78% de los sujetos de la muestra.